CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS
DEL INSTITUTO POLITÉCNICO NACIONAL
UNIDAD IRAPUATO

DESARROLLO DE HERRAMIENTAS BIOINFORMÁTICAS PARA ASISTIR LA PRODUCCIÓN HETERÓLOGA DE METABOLITOS SECUNDARIOS

Tesis que presenta

I.B.Q. Miguel Ángel Ramos Valdovinos

Para Obtener el Grado de

Maestro en Ciencias

En la Especialidad de Biotecnología de Plantas

Director de la Tesis

Dr. Agustino Martínez Antonio

Irapuato, Guanajuato Agosto de 2017

Esta tesis fue desarrollada en el Laboratorio de Ingeniería Biológica del Departamente de Ingeniería Genética del Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional Unidad Irapuato.

AGRADECIMIENTOS

Agradezco al Consejo Nacional de Ciencia y Tecnología por otorgarme una beca nacional para realizar mis estudios de Maestría.

Al Centro de Investigación y de Estudios Avanzados del Insituto Polítecnico Nacional Unidad Irapuato por prestar sus instalaciones para llevar a cabo esta tesis.

A mi director de tesis el Dr. Agustino Martínez Antonio, por su apoyo incondicional durante mi estancia en el Laboratorio de Ingeniería Biológica.

A mis tutores la Dra. Selene Lizbeth Fernández Valverde y el Dr. Cei Abreu Goodger por sus acertadas observarciones y sugerencias para llevar a cabo mi tesis.

Al grupo del Laboratorio de Ingeniería Biológica, especialmente a Paola por todo su apoyo.

Contenido

Índice de ecuaciones

Índice de figuras

Resumen

La sobreexplotación de recursos naturales y la alta demanda de insumos para las industrias motiva la búsqueda de nuevas fuentes de materias primas; entre ellas la utilización de microorganismos para producir compuestos de alto valor. Cuando no es viable utilizar a los microorganismos que producen naturalmente un compuesto, se recurre a la Biología Sintética para producirlos en otros organismos. Un metabolito se puede producir a través de cientos de rutas posibles. Por ello, en este trabajo desarrollamos un programa que es capaz de encontrar esas rutas metabólicas a partir de la información de la colección BioCyc. Las rutas son representadas como hipergrafos, partiendo de los compuestos producidos por un organismo. Para cada ruta implementamos una evaluación termodinámica que permite ordenarlas de acuerdo con la de las reacciones químicas limitantes y la similitud del precursor con el compuesto deseado utilizando la geometría Manhattan. Una vez seleccionada una ruta metabólica el programa sugiere genes de acuerdo con la información disponible en la base de datos, como son la evidencia experimental, el tamaño de genes, la cantidad de reacciones que realiza en la ruta, entre otras. Para continuar con el flujo de trabajo se brindan algunas herramientas para la visualización y edición de secuencias. Nuestra herramienta indica que se pueden producir hasta 5,618 metabolitos en más de 7,000 organismos y que la mayoría de esos compuestos se pueden producir introduciendo menos de 44 reacciones en cualquier organismo. Por ejemplo, para el organismo modelo *Escherichia coli* se puede extender su espacio metabólico hasta los 4,361 compuestos utilizando menos de 19 reacciones, esto representa un incremento del 40% comparado a lo reportado por Carbonell y colaboradores en 2012. El desarrollo de estas herramientas sienta las bases para la consideración de nuevos organismos en la producción de metabolitos de alto valor. Además, ahorra tiempo en el diseño de los proyectos ya que automatiza la búsqueda de rutas metabólicas que es uno de los primeros pasos de cada proyecto. Utilizando índices de centralidad, se encontraron los actores principales de un grafo creado a partir de las reacciones anotadas en BioCyc. Estos actores representan compuestos que participan en un gran número de rutas y cuya producción puede ser blanco de optimización en

cualquier organismo para aumentar la obtención de los compuestos deseados mediante ingeniería metabólica.

Abstract

The overexploitation of natural resources and the high demand of supplies for industries motivates the search for new sources and ways of producing raw materials. An alternative is the use of microorganisms to produce high-value compounds. When it is not feasible to use microorganisms that naturally produce a compound, Synthetic Biology is used to produce them in model organisms. A metabolite can be produced through hundreds of possible metabolic routes. Thus, in this work we developed a suite of programs that can identify those metabolic routes using the BioCyc collection as the main source of information. The routes are represented by hypergraphs, starting from the compounds produced by an organism. For each metabolic route, we implement a thermodynamic evaluation that allows us to order them according to the driving force of the limiting chemical reactions and the similarity of the precursor with the desired compound using the Manhattan geometry. Once a metabolic pathway is selected, the program suggests genes per the information available in the database; such as experimental evidence, size of genes, the number of reactions performed on the route, among others. To continue with the workflow, some tools are provided for viewing and editing sequences. As a result, we found that up to 5,618 metabolites can be produced in more than 7,000 organisms and that any of these compounds can be produced by introducing less than 44 reactions in the model organism *Escherichia coli*. That is, the metabolic space of *E. coli* can be extended to 4,361 compounds using less than 19 reactions, representing an increase of 40% of potential metabolites produced by *E. coli* compared to the reported by Carbonell and collaborators in 2012. The development of these tools sets may facilitate the incorporation of new organisms in the production of metabolites of high value, besides saving time as it automates the first steps during the design of a biotechnology project. Through the analysis of the extended metabolic network of BioCyc we find the central metabolites, which can be target of optimization in any organism to increase the production of any desired compound via metabolic engineering.

1. Introducción

1.1 Producción heteróloga de metabolitos de alto valor

Existe un interés creciente en la producción de compuestos de alto valor utilizando microorganismos, esto debido al potencial agotamiento del petróleo y otros recursos naturales y por el impacto que tienen en el ambiente las maneras tradicionales de producción basadas en la industria química (Draths & Frost, 1995). Sin embargo, para la mayoría de los casos es necesario realizar modificaciones genéticas en estos organismos para optimizar la producción o, más frecuentemente, introducir nuevas rutas metabólicas (Chatsurachai, Furusawa, & Shimizu, 2012).

Para la producción de metabolitos heterólogos son pocos los organismos modelo con los que se pueden trabajar. Esto a pesar de que existen organismos con reservorios de precursores más grandes, y que tienen mayor tolerancia a los compuestos intermediarios y finales o que tienen una mayor producción debido a una mayor robustez de su metabolismo interno con respecto a la bacteria normalmente usada que es, *Escherichia coli* (Verhoef, Ruijssenaars, de Bont, & Wery, 2007). Esta limitación se debe en parte a la carencia de herramientas experimentales como bioinformáticas que permitan trabajar con estos organismos no modelos. Por esta razón, se considera que poner a la disposición de los investigadores herramientas para trabajar con estos organismos no modelos puede incentivar la investigación y el desarrollo de una biotecnología más efectiva (Loescheke & Thies, 2015).

El proceso básico que se sigue para la producción de un metabolito en un organismo de forma heteróloga, se puede definir en los siguientes pasos:

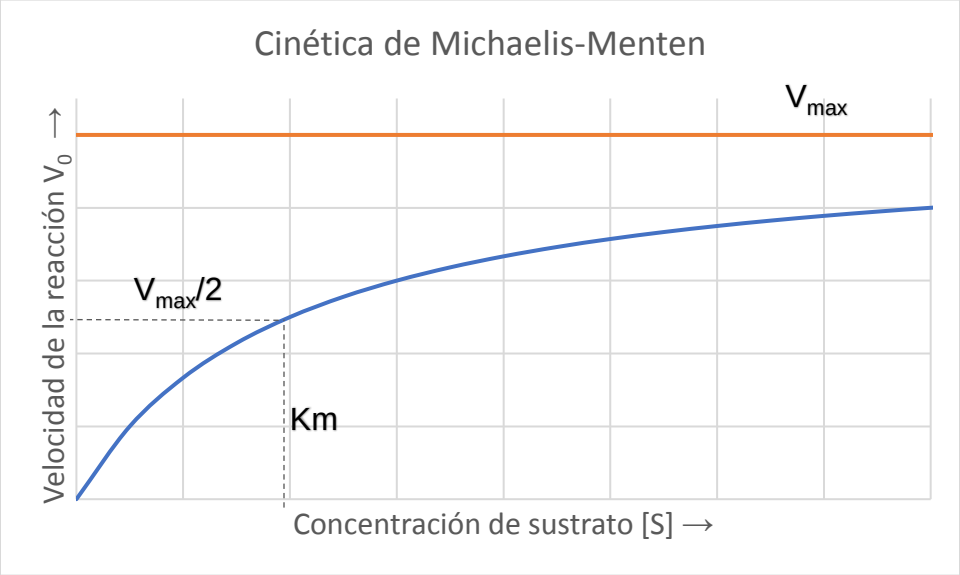
1. Selección del compuesto a producir.
2. Selección del organismo a utilizar.
3. Búsqueda y selección de rutas para producir de forma heteróloga el compuesto.
4. Búsqueda de las enzimas involucradas en la ruta seleccionada.
5. Selección de los genes que codifican para las enzimas.
6. Extracción de los genes a partir de otros organismos o por síntesis química.

7. Diseño de un vector para introducir el o los genes.
8. Introducción de los genes al organismo seleccionado.
9. Validar la producción.
10. Optimizar la producción.
11. Escalar la producción, etc.

Una vez asegurado que el organismo es capaz de producir un metabolito de forma heteróloga se continúa con un proceso de optimización que va desde adaptar el uso de codones de los genes al organismo, hasta hacer ingeniería metabólica en el organismo utilizado y finalmente un proceso de optimización de las condiciones de producción en biorreactores.

1.2 Propiedades cinéticas de las enzimas

Las enzimas son catalizadores biológicos cuya función es acelerar la velocidad a la cual ocurren las reacciones dentro de un organismo. La velocidad a la que una enzima puede llevar a cabo una reacción (V_0) depende de la concentración de su sustrato ([S]), y se define como la cantidad de moles de productos que se forman por segundo. El incremento en V_0 inicia siendo lineal con respecto a [S], pero conforme [S] aumenta se va reduciendo el incremento de V_0, hasta que deja de crecer. A este punto de la velocidad se le conoce como V_{max} y es la velocidad de reacción máxima que puede lograr una enzima. En la **Gráfica 1**, se muestra un parámetro llamado Km, que es la constante de Michaelis-Menten, la cual indica cual es la concentración de sustrato a la cual la velocidad de la reacción es la mitad de la velocidad máxima que puede lograr la enzima. Entre menor sea el valor de Km para una enzima, mayor será la afinidad que tenga la enzima por su sustrato y por tanto tendrá más eficiencia, en los organismos vivos se ha encontrado que la concentración de sustratos dentro de las células corresponde con el valor de Km (Berg, Tymoczko, & Stryer, 2002). Si se tiene una enzima que tiene una alta Km la única forma de aumentar la velocidad de la reacción para una concentración de sustrato fija es aumentando la concentración de la enzima, esto es difícil ya que represente un mayor consumo de recursos para las células de un organismo y en ocasiones no es posible debido a limitantes técnicas al sobrexpresar enzimas.



Gráfica 1. Principales parámetros en una cinética de Michaelis-Menten (Berg *et al*., 2002), gráfica modificada.
En esta gráfica se observa como el incremento en la V_0 va disminuyendo conforme aumenta [S], también se observa como V_0 es igual a $V_{max}/2$ para [S] igual a Km.

La mayoría de las reacciones pueden llevarse a cabo de forma reversible y esto ocurre simultáneamente. Cuando se inicia una reacción *in vitro* se parte de una concentración inicial de sustrato [S_0] la cual va disminuyendo conforme avanza el tiempo, al contrario del producto de la reacción que va aumentando. Para todas las reacciones llegará un momento en el cual la cantidad de reactivos y productos se mantendrán constantes, a este punto se le llama estado de equilibrio de la reacción y no significa que se haya detenido la reacción, sino que la velocidad a la cual se forman los productos es igual a la velocidad a la que se consumen (Berg *et al*., 2002).

1.3 Base de datos

1.3.1 Colección BioCyc

BioCyc es una colección de bases de datos de rutas metabólicas y genomas (PGDBs por las siglas en inglés de Pathway/Genome Databases). Esta colección

cuenta con un sitio web en el cual se encuentran disponibles de forma gratuita genomas y sus anotaciones para miles de organismos. Además de las bases de datos, ponen a disposición de los usuarios utilidades de entre las que sobresale el programa Pathway Tools que es capaz de reconstruir las rutas metabólicas de un organismo a partir de un genoma anotado siguiendo ciertas reglas (BioCyc no realiza anotaciones de genoma), y permite realizar correcciones manuales y visualizar la información generada. Así, por cada genoma agregado se crea una nueva PGDBs que recibe diferentes clasificaciones de acuerdo con la cantidad de curación manual que haya recibido (Karp, Latendresse, & Caspi, 2011). Por ejemplo, las PGDBs que han recibido un mayor número de curaciones y revisiones son las de *Escherichia coli* MG1655 (EcoCyc) y en MetaCyc lo que las clasifican como de nivel 1, mientras que las rutas que recibieron sólo un poco de curación manual son clasificadas como de nivel 2, y las que fueron creadas computacionalmente sin recibir curación manual son las de nivel 3. En estas últimas dos categorías se encuentran la mayoría de los PGDBs (BioCyc, 2016).

BioCyc tiene particularidades que la hacen sobresalir de las demás, por ejemplo, el libre acceso a toda su base de datos y la posibilidad de descargarla, además de que contiene anotaciones basadas en el formato SBML que facilitan el entendimiento de los procesos biológicos. La mayoría de las reacciones en BioCyc se encuentran balanceadas, lo que la hace una colección ideal para trabajar con modelos metabólicos (Altman, Travers, Kothari, Caspi, & Karp, 2013).

1.3.2 Enciclopedia de Genes y Genomas de Kioto (KEGG)

La Enciclopedia de Genes y Genomas de Kioto o KEGG por sus siglas en inglés, es un grupo de bases de datos que contiene información sobre organismos a diferentes niveles de organización. En ella se pueden encontrar genes obtenidos de genomas secuenciados con anotaciones sobre su papel biológico; esta información es almacenada en la base de datos GENES. En otro nivel llamado PATHWAY se encuentra la información sobre las rutas metabólicas reconstruidas a partir de la información de los genomas y de una base de datos llamada LIGAND en la cual se almacena información sobre compuestos químicos, enzimas y reacciones

enzimáticas. Esta base de datos está disponible desde su sitio web de forma libre, además tiene una Interfaz de Programación de Aplicaciones (API) basada en REST a través de la cual se pueden realizar consultas de forma automatizada. Esta base de datos, sin embargo, sólo está disponible para uso estrictamente académico, no permite el uso de la información para fines comerciales a diferencia de BioCyc que permite el uso comercial y académico de forma gratuita (Kanehisa & Goto, 2000).

Las bases de datos BioCyc y KEGG comparten muchas similitudes, por ejemplo, ambas toman genomas anotados (ninguna de las dos realiza anotaciones en genomas). Y a partir de estas anotaciones reconstruyen las posibles rutas metabólicas. También tienen herramientas que ponen a disposición de los usuarios en sus páginas web, y utilizan rutas de referencia curadas manualmente para producir las rutas a partir de los nuevos genomas anotados, entre otras. A pesar de estas semejanzas KEGG difiere de BioCyc en que mantiene una base de datos de compuestos sintéticos y de medicamentos, así como que requiere una licencia de paga para usar comercialmente la información de la base de datos (Altman, Travers, Kothari, Caspi, & Karp, 2013).

A pesar de que existen otras bases de datos de reacciones y rutas como Rhea, BiGG, UniPathway, BioPath, Reactome y BioCyc. Muchas de estas bases de datos fueron generadas a partir de BioCyc o KEGG o bien sólo tienen links a ellas (Altman *et al*., 2013).

1.3.3 Fuente Universal de Proteínas (Uniprot)

Uniprot es una iniciativa internacional para proveer a la comunidad científica con una base de datos en la cual se pueda centralizar la anotación de secuencias de proteínas y su información funcional, esta base de datos es actualizada cada 4 semanas (Boutet *et al*., 2016). La información de las anotaciones que han sido curadas manualmente está en la base de datos UniProtKB/Swiss-Prot y sirve de referencia para generar información sobre nuevos organismos en UniProtKB/TrEMBL utilizando algunas reglas. Uniprot está ligada a otras 140 bases de datos con la cuales mantiene referencias cruzadas y con algunas de las cuales

mantiene colaboración. Uno de los objetivos de la base de datos es unificar secuencias de proteínas, por lo que buscan eliminar redundancias en su base de datos eliminando secuencias iguales que provienen de la misma especia referenciándola a una secuencia consenso. En otros casos, se unen secuencias que tienen un alto grado de similitud y se hacen anotaciones sobre las posibles variantes que se pueden encontrar. La información completa de Uniprot puede ser descargada en formato de texto o se puede consultar directamente desde el sitio web, donde existe una mayor cantidad de formatos de descarga (UniProt Consortium, 2017).

1.4 Enfoques para la reconstrucción de rutas metabólicas

La identificación y enumeración de rutas metabólicas, aunque parece simple no es una tarea trivial, ya que enfrenta desafíos que requieren abordar el problema desde distintos puntos de vista, siendo la principal limitante la disponibilidad y confiabilidad de la información metabólica. Actualmente, existen cientos de bases de datos que contienen información diversa sobre componentes e interacciones biológicas y a pesar de que existen esfuerzos enormes por unificar la información, esto no se ha logrado del todo (Altman, Travers, Kothari, Caspi, & Karp, 2013). Además de que con los avances en las tecnologías de secuenciación la cantidad de información que se genera sobrepasa por mucho la velocidad a la cual se puede procesar. Por esta razón cada vez son más valiosos los programas que permiten automatizar procesos o generar nueva información. Las bases de datos KEGG (Kanehisa & Goto, 2000) y BioCyc (Caspi *et al*., 2016) permiten generar información sobre rutas metabólicas de organismos completos a partir de genomas con anotaciones y esta información está disponible libremente para ser leída por cualquier persona a través de sus páginas de internet. Además, ponen a la disposición de los usuarios distintas herramientas para visualizar, analizar o descargar la información. Una de las herramientas que ponen a disposición ambos sitios es la reconstrucción de rutas metabólicas de un compuesto a otro, estas herramientas llamadas PathComp en KEGG (Moriya *et al*., 2010) y Metabolic Route Search en BioCyc, generan rutas lineales que aumentan exponencialmente conforme a la cantidad de pasos máximos

introducidos. Cómo estas existen una gran cantidad de herramientas que permiten encontrar rutas metabólicas, y la complejidad de los algoritmos varía desde simplemente conectar reacciones hasta evaluar el impacto de una determinada ruta en el organismo utilizado. A pesar de la cantidad de herramientas que existen para la reconstrucción de rutas metabólicas la mayoría de ellas sólo permiten la reconstrucción de rutas de un compuesto a otro o que utilizan como chasis únicamente a *E. coli*, esto a pesar de que los algoritmos son diseñados para utilizarse en cualquier especie.

1.4.1 Aumento de la capacidad metabólica de los organismos

Al introducir nuevas reacciones en un organismo mediante ingeniería genética, se pueden producir nuevos metabolitos. El total de las reacciones bioquímicas que se pueden introducir en un organismo constituyen el espacio metabólico extendido de este, ya que extienden la cantidad de reacciones que puede llevar a cabo dicho organismo para producir compuestos (Chatsurachai, Furusawa, & Shimizu, 2012; Carbonell, Planson, Fichera & Faulon, 2011) . Esto puede ser generado a partir de la información existen en bases de datos como BioCyc o MetaCyc. Entre mayor sea el espacio metabólico extendido significa que es posible producir una mayor cantidad de metabolitos o que estos puedan ser producidos en menos pasos.

A partir de un compuesto es posible generar una gran cantidad de otros compuestos. Sin embargo, existen algunos compuestos que son claves en la expansión de las redes metabólicas, ya que si son eliminados reducen considerablemente el tamaño de la red metabólica. Que algunos compuestos tengan un papel crucial en la expansión de la red plantea la posibilidad de una historia evolutiva en la cual los organismos podían sólo producir unos cuantos compuestos y a partir de ellos se fueron expandiendo las redes metabólicas gracias a procesos evolutivos (Handorf, Ebenhöh, & Heinrich, 2005).

El número de diferentes enzimas conocidas es incierto en parte debido a la redundancia de información en las bases de datos (Altman, Travers, Kothari, Caspi,

& Karp, 2013). Sin embargo, el número de las reacciones están en el orden de las decenas de miles.

1.4.2 Retrosíntesis

La retrosíntesis es utilizada ampliamente en la síntesis química de compuestos, en la cual se consideran transformaciones químicas conocidas de forma inversa para conocer los precursores de un compuesto. Esto se repite hasta que se llega a un compuesto que está ampliamente disponible o que es muy económico (Carbonell, Planson, Fichera, & Faulon, 2011). La retrosíntesis en redes metabólicas es muy parecida, sólo es cuestión de cambiar las reacciones químicas por reacciones enzimáticas y los compuestos precursores por compuestos producidos por un organismo o alguno que puede ser añadido al medio de cultivo.

Las transformaciones químicas para la retrosíntesis en rutas metabólicas son obtenidas de bases de datos de enzimas en las cuales cada enzima realiza una transformación de un compuesto en otro. Es posible hacer una abstracción de la actividad de las enzimas e ir más allá de la transformación de un compuesto a otro, pensando en cada transformación como en una operación que puede ser aplicada a cualquier compuesto que reúna ciertas estructuras químicas (Li *et al*., 2004). Con esto es posible predecir qué enzimas pudieran llevar a cabo una reacción que no es conocida o incluso llevar a cabo experimentos de evolución dirigida para crear enzimas con nuevas actividades catalíticas. Llevar a cabo este tipo de predicciones es muy complejo y computacionalmente tardado, por lo que en la mayoría de los casos es reservado para generar nuevas rutas a compuestos que no se conoce cómo se sintetizan.

Para buscar rutas de los compuestos que produce un organismo a un compuesto deseado, la retrosíntesis es más rápida computacionalmente, debido a que las rutas se expanden desde el compuesto deseado hasta llegar a algún compuesto del organismo, en lugar de generar rutas para cada compuesto del organismo hasta llegar el compuesto deseado (Carbonell, Fichera, Pandit, & Faulon, 2012).

La cantidad de combinaciones de reacciones durante la retrosíntesis se vuelve tan grande durante la búsqueda de rutas metabólicas que se han desarrollado algoritmos heurísticos en los cuales se determina si una reacción acerca o aleja una ruta del compuesto deseado, en la **Tabla 1** se da un ejemplo para que el lector dimensione la cantidad de rutas lineales que se pueden generar. Un ejemplo de esto es el trabajo realizado por McShan y colaboradores en 2003, en el cual utilizan las geometrías Manhattan y Euclidiana para determinar si una reacción producía un compuesto más parecido al compuesto buscado o lo hacía más diferente de acuerdo con su fórmula química, con esta información se podía guiar al algoritmo para evitar las reacciones que no condujeran a ningún resultado desde antes de simularlas, ahorrando tiempo de cómputo.

Tabla 1. Parámetros en la enumeración de rutas para la producción de all-*trans*-fitoeno a partir de *E. coli*.
Esta tabla da un ejemplo de la cantidad de rutas lineales que se pueden generar durante la retrosíntesis. No todas ellas son viables.

Máximo de reacciones	Pares reactantes	Compuestos conectados	Rutas generadas*	Hiperrutas
1	4	6	2	0
2	10	10	6	0
3	21	21	31	1
4	31	31	172	3
5	39	39	391	7
6	44	44	549	12
7	51	51	1,075	16
8	54	54	3,827	18
9	57	57	13,649	18
10	60	60	37,029	18
11	62	62	389,524	18

*Incluye rutas lineales inconclusas que no ligan a un compuesto que produce el organismo con el compuesto deseado, o rutas que predicen la formación de los compuestos deseado a partir de un compuesto ubicuo como el agua.

1.4.3 Redes metabólicas como grafos

Las redes son una representación informal para elementos que están conectados o que interaccionan entre ellos. Las redes pueden ser representadas como grafos donde cada elemento de la red se le denomina vértice o nodo, las interacciones

entre dos nodos son llamadas aristas y el conjunto de aristas que se deben cruzar para ligar un nodo a otro se le llama ruta y para este trabajo representan rutas metabólicas (Junker & Schreiber, 2008). Las redes pueden tener dirección, dependiendo de las propiedades de la arista, por ejemplo, se puede hablar de una red metabólica dirigida si algunas de las reacciones sólo se pueden dar en un solo sentido, como es en el caso de las reacciones irreversibles, mientras que si la interacción ocurre en ambos sentidos o no se conoce la dirección se usan las redes no dirigidas. Para representar una reacción reversible en una reacción dirigida se pueden usar dos aristas con dirección contraria, sin embargo, esto hace que el grafo tenga aristas paralelas o aristas múltiples, lo que lo convierte en un grafo complejo o multígrafo y debe tratarse de forma distinta a los grafos simples (Hagberg, Schult, & Swart, 2008). El uso de redes metabólicas dirigidas es importante por la presencia de rutas que sólo pueden ocurrir en una dirección debido a que algunas de sus reacciones son irreversibles, además tienen la ventaja de que su simulación es menos demandante computacionalmente que para las redes no dirigidas.

Un par de ejemplos de herramientas para la enumeración de rutas metabólicas que se basan en grafos son PathComp de KEGG (Moriya *et al*., 2010) y Metabolic Route Search de BioCyc. Estas herramientas utilizan alineamientos de reactivos y productos a los que KEGG llama pares reactantes. Estos pares reactantes son generados manualmente en la base de datos KEGG debido a que existen algunos compuestos que no cumplen un papel principal en las reacciones, por ejemplo, los cofactores o intercambiadores de grupos funcionales, y deben ser ignorados para mejorar las predicciones (Moriya *et al*., 2010). Mientras que en BioCyc estos pares reactantes son generados por la combinación de reactivos con productos, lo cual produce errores al crear pares reactantes sin significado biológico, ya que sugiere la producción de compuestos a partir de otros compuestos altamente conectados. Un ejemplo se muestra en la **Figura 1,** donde se indica que el piruvato se puede convertir en agua y esta a su vez da origen a ATP en las reacciones de producción de geranil difosfato. Esto es incorrecto, ya que, si bien en las reacciones ocurren esos pasos, no es posible producir ATP a partir de una sola molécula de agua. Errores como estos son admitidos en la base de datos debido a que no existe una

forma automática de crear los pares reactantes, porque en algunos casos una molécula altamente conectada puede funcionar como transportador de grupos funcionales, pero en otros puede servir como precursor de algún compuesto. Este es el caso del ATP que sirve como moneda de cambio energético en muchas reacciones, sin embargo, también es el precursor de la 5'-5'''-diadenosina trifosfato. La curación manual de estos pares reactantes requiere demasiado trabajo humano, además de que el proceso es arbitrario porque no existen definiciones claras para algunas reacciones acerca de cuándo un compuesto es un precursor o sólo es un transportador de grupos funcionales para ciertas reacciones.

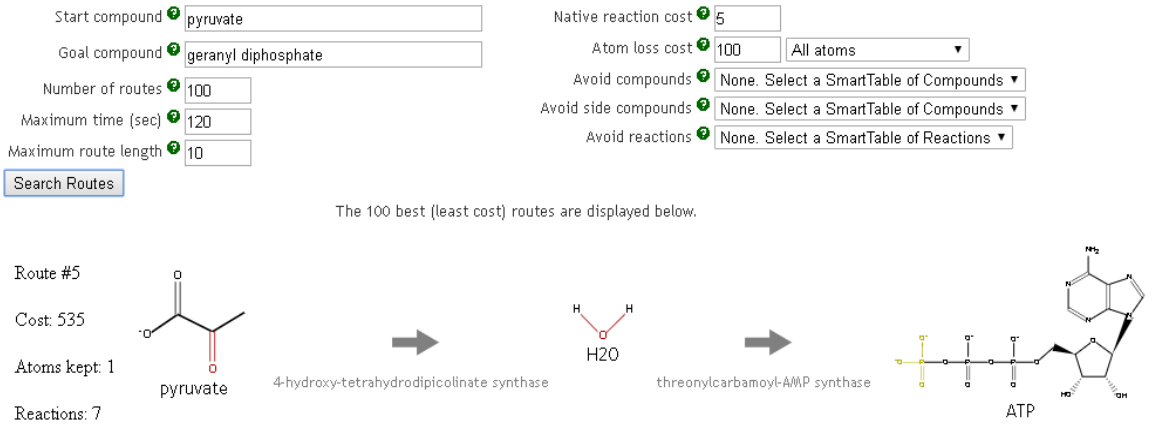


Figura 1. Reconstrucción de rutas con pasos carentes de significado biológico, imagen recortada *(SRI International, 2017)*.

En esta imagen se muestra como el algoritmo de Metabolic Route Search predice la formación de ATP a partir de agua, lo cual es incorrecto. Esto error se genera por el uso de pares reactantes generados computacionalmente, para crear rutas metabólicas lineales.

El número de rutas encontradas en los grafos suele seguir un crecimiento tipo exponencial de acuerdo con la cantidad de aristas permitidas (Newman, 2001). Debido a esto, las consultas se pueden generar cientos de miles de rutas desde un compuesto a otro, en especial si los compuestos intermediarios son precursores de muchos compuestos. La cantidad de rutas generadas se vuelve un problema más que un avance en la producción de metabolitos heterólogos, en especial si algunas de esas rutas no son viables. Esto sucede cuando al menos una reacción en la ruta requiere de más de un precursor, lo que sucede en el 83% de los casos (dato de

este trabajo) y estos precursores no siempre están disponibles en el organismo utilizado ni se sintetiza en la misma ruta.

1.4.4 Índices de centralidad en grafos

Los índices de centralidad son valores utilizados para analizar redes sociales, estos valores califican la importancia de los actores dentro de una red basados en sus interacciones y la mayoría se calculan por la presencia de los nodos dentro de las rutas más cortas. Debido a la implementación de redes en otras áreas, también se ha encontrado que se pueden inferir nodos importantes para otros tipos de redes a partir de su índice de centralidad (Brandes, 2001).

Centralidad intermedia

La centralidad intermedia de un nodo v es la suma de la fracción de las rutas más cortas para todos los pares de nodos que pasan por v, ver **Ecuación 1** (NetworkX Developers, 2015).

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \qquad \textbf{Ecuación 1}$$

Donde V son los nodos de la red, $\sigma(s,t)$ es el número de rutas más cortas que conectan los compuestos s y t, $\sigma(s,t|v)$ es el número de rutas de $\sigma(s,t)$ que pasan por el nodo v.

La centralidad intermedia es muy parecida a la centralidad de la carga, sin embargo, la centralidad de la carga es un índice que da una visión más global de la red, mientras que la centralidad intermedia suele dar una representación más local.

Centralidad de cercanía

La centralidad de cercanía de un nodo u es el recíproco de la suma de la distancia de las rutas más cortas de u a todos los v-1 nodos. Para normalizar esta función se multiplica por (n – 1), ver **Ecuación 2** (NetworkX Developers, 2015).

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$ **Ecuación 2**

Donde d(v,u) es la ruta más corta entre v y u, n es el número de nodos de la red.

Este índice busca representar la distancia que existe de un nodo cualquiera al resto de los nodos de una red. Entre más alto sea el valor de este índice de centralidad, significa que mayor es la cercanía de este nodo con todos los nodos a los que está ligado.

Centralidad de la carga

Según la definición de NetworkX, la centralidad de la carga de un nodo es la fracción de todas las rutas más cortas de la red que pasan a través de ese nodo, de forma genérica se puede calcular con la **Ecuación 3** (NetworkX Developers, 2015).

$$b = \sum_{s,t \in V} \sigma(s,t|v)$$ **Ecuación 3**

Donde σ(s,t|v) es la cantidad de rutas más cortas entre s y t que pasan por v. Esta ecuación puede ser normalizada aplicando b=b/(n-1)(n-2) para que la escala sea de 0 a 1.

Cuando los grafos contienen pares de nodos unidos por más de una arista se les llama multígrafos. Los multígrafos suelen producir errores con algunos algoritmos para calcular la centralidad de la carga debido a que se producen bucles en la búsqueda de las rutas más cortas, por esta razón existen algoritmos que utilizan propiedades internas de los grafos para calcular la distribución de la carga (Goh, Kahng, & Kim, 2001).

Esta centralidad también puede ser llamada por algunos autores o programas como centralidad del estrés.

1.4.5 Redes metabólicas como hipergrafos

Cuando un grafo tiene aristas que unen más de dos nodos, se les conoce como hipergrafos y a las aristas hiperarcos, mientras que a las rutas se les llama

hiperrutas (Carbonell, Fichera, Pandit, & Faulon, 2012). Simular redes metabólicas con hipergrafos permite eliminar algunos de los problemas encontrados al utilizar grafos. Por ejemplo, las reacciones enzimáticas son mejor representadas mediante hiperarcos ya que permiten tomar en cuenta todos los reactivos y productos presentes en la reacción. Cuando hay rutas con reacciones que requieren de más de un precursor para llevarse a cabo se pueden tomar en cuenta gracias a las hiperrutas ya que estas permiten simular la producción de un metabolito a partir de un conjunto de precursores. Además, el número de rutas se ve disminuido ya que varias rutas en un grafo pueden dar lugar a una ruta única en un hipergrafo y se eliminan las rutas que no son viables, esto se puede observar en la **Tabla 1** donde la cantidad de rutas lineales es menor a la de hiperrutas. Aún con todas estas mejoras, las hiperrutas no están exentas de errores que también son encontrados en las rutas, como puede ser la redundancia en la producción de compuestos o en el uso de reacciones, por lo que Carbonell y colaboradores utilizan el concepto de hiperrutas mínimas en las cuales, si se elimina cualquier reacción la hiperruta se ve interrumpida (Carbonell *et al*., 2012).

1.5 Evaluación y ordenamiento de las rutas metabólicas

Cuando la enumeración de rutas metabólicas genera demasiados resultados, no resulta de gran utilidad, esto debido a que siempre se busca la ruta que pueda dar una mayor producción y hay que realizar la evaluación de cada una de ellas para poder determinar cuál es la que más conviene para llevar a cabo un proyecto. Existen varios métodos para la evaluación de rutas metabólicas, por ejemplo, Carbonell y colaboradores en 2014 desarrollaron una plataforma llamada XTMS en la cual utilizan información de varias bases de datos como KEGG, EcoCyc, RHEA y BRENDA para construir un espacio metabólico extendido a partir del cual enumeran hiperrutas para *E. coli*, y las ordenan utilizando el análisis de balance de flujo, toxicidad de los metabolitos intermedios, termodinámica de las reacciones, y eficiencia enzimática, entre otras consideraciones, buscando dar el mayor rendimiento posible en la producción (Carbonel *et al*., 2014). Estas herramientas a pesar de que fueron pensadas para poder trabajar con cualquier organismo, sólo

han podido ser liberadas para *E. coli* debido a que es el organismo mejor estudiado. Tal es el caso que si se quisiera escalar la plataforma XTMS a otro organismo se encontrarían con la carencia de modelos para el balance de flujo y los valores de toxicidad de los intermediarios ya que los utilizados por ellos fueron calculados para *E. coli*.

1.5.1 Evaluación termodinámica de las rutas metabólicas

Las leyes de la termodinámica pueden ser aplicadas para mejorar el entendimiento de cómo suceden las reacciones químicas y las rutas metabólicas dentro de los organismos. Con un análisis termodinámico se puede estimar la velocidad a la cual ocurren las reacciones, la concentración de reactivos que se requieren para que sean posibles, su grado de reversibilidad, entre otros parámetros (Noor *et al*., 2014). El Cambio en la Energía Libre Estándar de Gibbs para una Reacción ($\Delta_r G'°$) es utilizado para la reconstrucción de rutas metabólicas. Sin embargo, las reacciones para las que ha sido calculado experimentalmente son muy pocas. Por ejemplo, para *E. coli*, que es un organismo muy estudiado, sólo tiene información experimental para el 8.1% de sus reacciones (Jankowski, Henry, Broadbelt, & Hatzimanikatis, 2008). Por esta razón se han desarrollado métodos de estimación para llenar los vacíos restantes, uno de los cuales es el método de contribución de grupo. Este método ha sido actualizado y adaptado a distintas aplicaciones, entre las que se encuentran los sistemas biológicos. A continuación, se resumirán brevemente las mejoras realizadas al método por Jankowski y colaboradores en 2008, y Noor y colaboradores en 2013.

Cálculo del $\Delta_r G'°$ a partir de la energía libre de formación de los reactivos

Se puede determinar el cambio en la energía libre estándar de Gibbs para la formación de un compuesto ($\Delta_f G'°$) mediante medidas experimentales, y estos a su vez pueden ser utilizados para calcular el $\Delta_r G'°$ para una reacción en la que estén involucrados, utilizando la **Ecuación 4**.

$$\Delta_r G'^{\circ} = \sum_{j=1} b_j * \Delta_f G'^{\circ}{}_j - \sum_{i=1} a_i * \Delta_f G'^{\circ}{}_i$$ **Ecuación 4**

Donde

$\Delta_f G'^{\circ}{}_i$ Cambio en la energía libre estándar de formación de un reactivo i

$\Delta_f G'^{\circ}{}_j$ Cambio en la energía libre estándar de formación de un producto j

a Coeficiente estequiométrico del reactivo i

b Coeficiente estequiométrico del producto j

Cuando se conoce el valor del $\Delta_r G'^{\circ}$ y sólo se desconoce el valor del $\Delta_f G'^{\circ}$ de uno de los reactivos, se puede calcular despejándolo de la formula y este reactivo puede entonces ser utilizado para encontrar otros valores de $\Delta_r G'^{\circ}$.

Estimación de la energía libre de reacción a partir del método de contribución de grupos

Para calcular el $\Delta_f G'^{\circ}$ de un reactivo se puede descomponer en grupos de átomos en los que cada uno de ellos tiene una contribución en el $\Delta_f G'^{\circ}$ del compuesto, ver **Ecuación 5**. Sin embargo, esta suposición no es siempre correcta debido a que existe también una contribución provocada por la proximidad de ciertos grupos funcionales y las combinaciones de ellos. Por esta razón, también existen valores que se deben añadir para tomar en cuenta el efecto acumulativo de ciertos grupos funcionales o enlaces, estos valores suelen ser encontrados en tablas; Jankowski y colaboradores en 2008 publicaron una lista de grupos de átomos y sus valores (Jankowski *et al*., 2008).

$$\Delta_f G'^{\circ}{}_{est} = \sum_{i=1}^{N_{gr}} n_i \Delta_{gr} G'^{\circ}{}_i$$ **Ecuación 5**

Donde

$\Delta_f G'^\circ_{est}$ Es el $\Delta_f G'^\circ$ estimado

N_{gr} Es el número de grupos conocidos para el compuesto

n_i Es el número de veces que está presente el grupo en el compuesto

$\Delta_{gr} G'^\circ_i$ Es la contribución del grupo i

Con los $\Delta_f G'$ encontrados se pueden calcular nuevas $\Delta_r G'^\circ$, pero una de las mejoras en el método de contribución de grupos es que no es necesario calcular todos los grupos, ya que cuando se calcula un $\Delta_r G'^\circ$ se utiliza la **Ecuación 6**.

$$\Delta_r G'^\circ_{est} = \sum_{i=1}^{m} v_i \left(\sum_{j=1}^{N_{gr}} n_j \Delta_{gr} G'^\circ_j \right) \qquad \textbf{Ecuación 6}$$

Donde

$\Delta_r G'^\circ_{est}$ Es el $\Delta_r G'^\circ$ estimado

v_i Es el coeficiente estequiométrico del reactivo i en la reacción

n_j Es las veces que está presente el grupo en un compuesto

$\Delta_{gr} G'^\circ_j$ Es la contribución del grupo j

N_{gr} Es el número de grupos para el compuesto

m Es el coeficiente estequiométrico de los compuestos involucrados en la reacción, siendo negativo para los reactivos y positivo para los productos.

En la **Ecuación 6** para muchas reacciones, si existen grupos desconocidos entre los reactivos y están presentes en ambos lados de la ecuación, estos grupos se cancelan y el $\Delta_r G'^\circ_{est}$ es sólo afectado por los grupos cambiantes. Esto permite calcular $\Delta_r G'^\circ_{est}$ incluso para reacciones con valores desconocidos de $\Delta_f G'^\circ$.

Cómo estos valores no son siempre exactos, se busca obtener los resultados que varíen lo menos posible. Por esta razón se hace uso de algoritmos de ajustes, en los cuales se introducen problemas cuyo resultado se conoce para entrenar al algoritmo y de acuerdo con esos valores el algoritmo va ajustando los resultados para que se ajusten lo más posible con los datos generados experimentalmente.

Programación lineal

La posibilidad de que una ruta metabólica ocurra en un sentido determinado depende de la concentración de los metabolitos que intervienen en ella. Encontrar las concentraciones de compuestos que permitan que la ruta vaya más rápido en la dirección deseada, se convierte en un problema de optimización en el cual se quiere tener la velocidad máxima de las reacciones de la ruta en función de la concentración de metabolitos, estos problemas pueden ser resueltos de forma más o menos sencilla utilizando programación lineal. Para que un problema de optimización lineal se pueda resolver necesita ser representado con la forma de la **Ecuación 7**.

$$f = c_1 x_1 + c_2 x_2 \cdots c_n x_n$$ **Ecuación 7**

Estas funciones son denominadas ecuaciones lineales y el número de soluciones que pueden tener van desde cero hasta una infinidad. En algunos casos sólo una de esas soluciones es la buscada y esta solución puede ser la que dé el mayor o menor resultado posible al evaluar la función, por lo que el problema de encontrar una solución se convierte en un problema de optimización lineal. La ecuación a la cual se quiere optimizar se le llama función objetivo y se pueden asignar ciertas limitantes a los valores que pueden tomar sus variables.

Ejemplo de problema lineal con restricciones

Maximizar	Z = X + 3Y - Z
Sujeto a	X + Y = 5
	Z – X = 4

Para facilitar la resolución de los problemas lineales, estos se consideran como problemas de maximización. Si una función requiere ser minimizada simplemente se considera maximizar esa misma función, pero invirtiendo el signo de toda la ecuación.

Las restricciones que se asignen a las variables pueden ser desigualdades, sin embargo, para introducirlas al programa lineal estas deben ser igualdades y se convierten añadiendo variables de decisión que deben cumplir con la condición de no ser negativas. Estas variables son llamadas variables de holgura si es una cantidad que hace falta para alcanzar un óptimo o variable de exceso si es una cantidad que sobrepasa el óptimo y son usadas en la siguiente forma:

Tabla 2. Ejemplos de la conversión de desigualdades en igualdades.

Para crear una igualdad a partir de una desigualdad es necesario sumar una variable de decisión, que puede ser una variable de holgura o exceso, dependiendo del tipo de desigualdad.

	Desigualdad	**Igualdad**
Variable de holgura	X + Y <= 50	$X + Y + V_h = 50$
Variable de exceso	X + Y >= 20	$X + Y – V_e = 20$

Los problemas lineales tienen una forma estándar, la cual es necesaria para resolverlos, la cual consiste en convertir el problema en un problema de maximización y eliminar las desigualdades de las restricciones.

Así un problema como el siguiente puede ser convertido a la forma estándar.

Forma no estándar.

Minimizar W = - X - 2Y + 4Z
Restricciones X + Z >= 4
Z – Y <= 5

Forma estándar.

Maximizar X + 2Y - 4Z = 0
Sujeto a X + Z - V1 = 4
- Y + Z + V2 = 5
X, Y, Z, V1, V2 >= 0

Existen variables sin restricciones que pueden tener valores negativos, sin embargo, aquí no se expondrá cómo introducirlas debido a que no son necesarias para este trabajo.

Con el programa lineal en su forma estándar se siguen los siguientes pasos de forma iterativa, hasta encontrar una solución óptima o un error.

1. Si todas las variables de la primera fila en forma estándar no son negativas, se ha encontrado la solución óptima, y se terminan las iteraciones.
2. Se toma la columna con el valor más negativo en la primera fila y esta variable ahora se considera una variable entrante.
3. Para cada una de las filas excepto la primera, se hace la división de la última columna de la derecha entre el valor de esa misma fila en la columna de la variable entrante y se selecciona la que dé el valor menor no negativo de entre todas las filas para llamarla la variable saliente o eje pivotante.
4. Si no se encuentra algún valor en el paso anterior se considera que el programa lineal no tiene límites y por tanto no existe un máximo, a pesar de que existe infinidad de soluciones.
5. Se aplica la eliminación de Gauss-Jordan en el eje del pivote.
6. Se repite el proceso desde el paso 1.

Si no existen soluciones para el problema lineal, se dice que es imposible, mientras que si la solución siempre está creciendo se dice que el problema no tiene límites y por tanto no existe un valor óptimo. Estos son los dos principales problemas que se pueden encontrar al momento de optimizar una ecuación, si no ocurre ninguna de las dos suposiciones anteriores se dice que el problema sí tiene un óptimo (Larson & Falvo, 2017).

2. Objetivos

2.1 Objetivo General

- Desarrollar herramientas informáticas para asistir la producción heteróloga de metabolitos.

2.2 Objetivos Específicos

- Crear un algoritmo que identifique todas las posibles rutas metabólicas que desde un organismo modelo conecten hasta un compuesto heterólogo deseado y retorne todos los genes heterólogos involucrados.
- Ordenar las rutas encontradas de acuerdo con los valores termodinámicos de las reacciones involucradas y al costo en biomasa.
- Compilar una base de datos de los organismos y reacciones enzimáticas para buscar las rutas metabólicas.
- Desarrollar una interfaz de usuario para la edición de genes y su ensamblado *in silico*.

3. Materiales y Métodos

3.1 Equipos

Para la realización del proyecto se utilizó una laptop con 6 Gb de memoria RAM DDR3, procesador Intel® Core™ i5-3317U, con Windows 8 de 64 bits y 1 Tb de disco duro. A pesar de que los requisitos del programa son modestos, si se quiere compilar la base de datos completa se requiere de al menos 800 Gb de espacio libre en el disco duro, además de que para utilizar el programa es necesario 1.5 Gb de espacio libre, esto debido al peso de la base de datos. Por lo demás se recomienda al menos un procesador i5 o algún otro equivalente y al menos 2 Gb de RAM.

3.2 Programas

Los programas desarrollados en este proyecto fueron escritos principalmente en Java 8, utilizando el entorno de desarrollo NetBeans 8. Además de las librerías que vienen por defecto con java se utilizaron las librerías de la **Tabla 3**. La información de la base de datos se descomprimió y se procesó para generar Archivos Separados por Tabuladores (extensión *.TSV por las siglas en inglés de *Tab-Separed Values*), con estos archivos se creó una base de datos en SQL utilizando phpMyAdmin que viene incluido en XAMPP. Como la base de datos es muy grande el programa puede conectarse mediante PHP a un servidor SQL conteniéndola, por lo que la conexión entre el programa y la base de datos se realiza mediante el método POST de PHP.

Tabla 3. Librerías Java de terceros

Librería	Descripción	Enlace
Balloontip	Cuadros de información personalizados	http://timmolderez.be/balloontip
Commons Math3	Algoritmos matemáticos como optimización lineal	http://commons.apache.org/proper/commons-math/
Json Simple	Permite trabajar con objetos en formato json	https://code.google.com/archive/p/json-simple/
Jsoup	Analiza texto en formato html	https://jsoup.org/

Para procesar la red metabólica de BioCyc se utilizaron los programas Cytoscape 3.5 y NetworkX 1.11. NetworkX es un paquete escrito en Python, por lo que algunas dependencias tienen que ser instaladas de acuerdo con los requisitos del sistema operativo y el paquete, para consultar las páginas de los programas utilizados consultar la **Tabla 4**.

Tabla 4. Programas utilizados para este trabajo.

Programa	Enlace
Java	https://www.java.com
NetBeans	https://netbeans.org
NetworkX	https://networkx.github.io
XAMPP	https://www.apachefriends.org/es/index.html
Cytoscape	http://www.cytoscape.org

Utilizando NetworkX se calcularon los índices de centralidad, mientras que con Cytoscape se calcularon la distribución de la distancia de las rutas más cortas, el número de componentes de la red y el diámetro de la red. Esto se describe con más detalle en la **Sección 3.8**.

3.3 Generación de la base de datos local

Las bases de datos utilizadas fueron BioCyc, KEGG y Uniprot, aunque BioCyc fue en la que se basó principalmente este proyecto. De BioCyc se obtuvieron los organismos, los compuestos que tiene cada organismo, las reacciones conocidas que pueden utilizarse, las enzimas que realizan esas reacciones, los genes que codifican para las enzimas y los valores termodinámicos para las reacciones y compuestos. Mientras que de KEGG se obtuvieron genes no presentes en BioCyc que codifican para enzimas que realizan reacciones anotadas en BioCyc. De Uniprot se obtuvieron los valores para las constantes de Michaelis-Menten. Con la información anterior se generon las rutas metabólicas y se ordenaron utilizando sus propiedades termodinámicas y la similitud con el compuesto precursor, después se seleccionaron los genes necesarios para llevarlas a cabo, ver **Figura *2***.

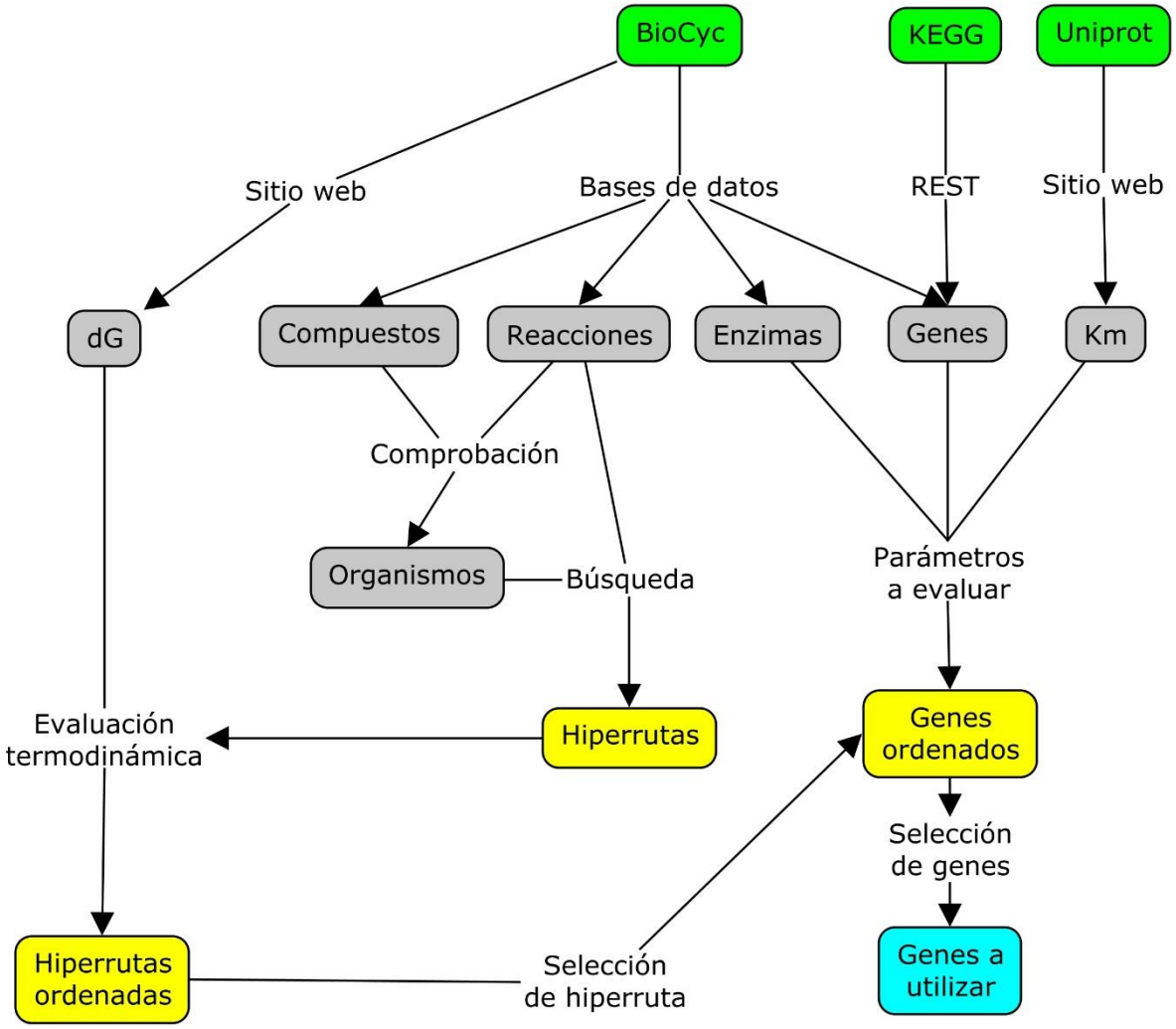


Figura 2. Flujo de información para generar las rutas metabólicas.

Las bases de datos se muestran en verde, la información recuperada en gris, los procesos del programa en amarillo y en azul el resultado. La información es recuperada de las bases de datos y almacenadas localmente para después ser utilizadas en la reconstrucción de rutas metabólicas.

3.3.1 Acceso a las bases de datos

Para trabajar con la colección de BioCyc se descargó la base de datos en formato de texto plano para lo cual se solicitó una licencia tipo "*Data File License*", desde el sitio oficial de BioCyc. Esta licencia es gratuita para uso académico y comercial. La

base de datos con la que se trabajó fue descargada el 20 de febrero de 2017 y los 7,624 organismos contenidos pesaban 74.5 Gb comprimidos.

También se utilizó la base de datos Uniprot, pero esta fue consultada directamente desde su sitio web, los identificadores de Uniprot presentes en BioCyc se consultaron y se descargaron todas sus anotaciones, las cuales se analizaron para tomar los valores de Km para cada enzima.

3.3.2 Recopilación de los compuestos

Se buscaron todos los compuestos presentes en BioCyc a partir de los archivos de texto plano, de los cuales se recuperó la información principal. Además, para cada uno de los compuestos se recuperó desde la página web de BioCyc el valor de la energía libre de formación, en caso de no encontrarse el valor fue establecido como 0 Kcal/mol.

Algunos compuestos genéricos no anotados en alguna base de datos tuvieron que ser creados para darle continuidad a algunas reacciones. Un ejemplo claro de estos compuestos fue NAD(P)H, identificado como NADH-P-OR-NOP en BioCyc, este compuesto genérico hace referencia a NADPH o NADH. La incorporación de estos compuestos permitió aumentar la reconstrucción de rutas de prueba de un 16% a un 77%. Para cada uno de los organismos creados se anotó que eran capaces de producir todos los compuestos genéricos del **Apéndice 2.**

3.3.3 Procesamiento de las reacciones

Para cada una de las reacciones presentes en los archivos de texto plano de BioCyc se recuperó la información más importante. Después se le asignó un valor único a cada reacción y se descargó el valor de la energía libre de reacción, si este valor no estaba disponible se calculó a partir de la energía libre de formación de sus reactivos y productos.

Algunos compuestos estaban anotados de forma genérica en BioCyc, por ejemplo, el compuesto llamado NAD(P) que en realidad no es un compuesto sino un grupo de compuestos formado por NAD y NADP. Este tipo de compuestos fueron

agregados manualmente a la base de datos para aumentar el número de reacciones posibles, como se menciona en el apartado anterior.

Las reacciones que eran capaces de llevarse a cabo de forma espontánea se guardaron como todas las demás reacciones. Además, se creó una lista con ellas, para ser utilizadas en la reconstrucción de rutas metabólicas.

BioCyc contiene reacciones compuestas que son reacciones formadas por otras sub-reacciones (ver **Figura 3**), dividir las reacciones permite describir reacciones de forma más detallada cuando se conocen los mecanismos de reacción, además estas reacciones tienen propiedades que les permite parecerse más a una ruta metabólica que a una reacción porque están formadas por secuencias de reacciones. Las reacciones compuestas se resumen con todos los compuestos que entran y salen de una reacción, sin tomar en cuenta los intermediarios. Las reacciones compuestas permiten separar sub-reacciones que se llevan de forma espontánea y por tanto no catalizadas por las enzimas o también permite separar reacciones en las cuales los intermediarios reaccionan sólo con algunas sub-unidades de complejos enzimáticos (SRI International, 2016). Estas reacciones fueron recuperadas y guardadas en un archivo para usarse evitando la reconstrucción de rutas redundantes.

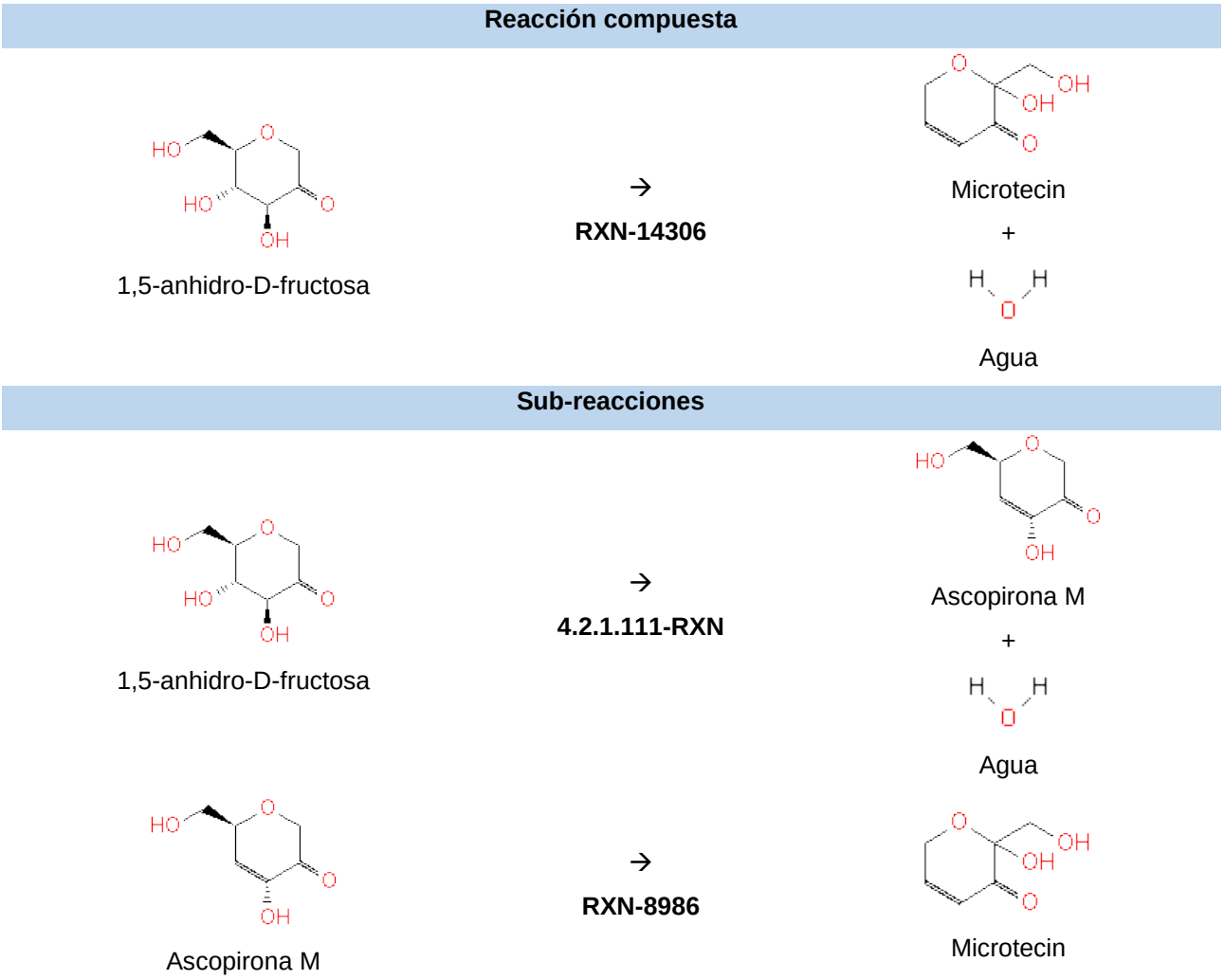


Figura 3. Ejemplo de una reacción compuesta.

La reacción compuesta RXN-14306 de BioCyc puede ser dividida en las sub-reacciones 4.2.1.111-RXN y RXN-8986, para dar una mejor explicación de cómo ocurre el mecanismo de reacción.

Por último, si una reacción unidireccional en un organismo se encuentraba en otro organismo con dirección contraria, se consideró que dicha reacción es bidireccional.

3.3.4 Compuestos producidos por organismos

La colección BioCyc maneja los elementos de su base de datos como lo hacen los lenguajes de programación orientada a objetos, utilizando clases que después se instancian a objetos, teniendo BioCyc un reservorio de clases (entidades biológicas) que pueden existir en cualquiera de sus bases de datos (organismos). Cuando se

quiere agregar un compuesto a una nueva base de datos, se crea una instancia de la clase del compuesto en la nueva base de datos, esta instancia se puede modificar para agregar anotaciones o propiedades y así describir mejor el contexto del organismo (Kerp *et al*., 2005).

Cada una de las clases está clasificada jerárquicamente y puede contener o ser contenida dentro de otras clases, en la **Figura 4** se muestra un ejemplo. Cuando se agrega a un organismo un compuesto, se crea una instancia de la clase que representa al compuesto y también de las subclases que contiene esa clase, a menos que se indique lo contrario (Kerp *et al*., 2005). Las clases que agrupan a varios compuestos representan compuestos genéricos que comparten alguna característica química o biológica en común, por ejemplo, el compuesto genérico "an alcohol" engloba a todos los compuestos que tienen un grupo funcional hidroxilo (OH). Esta clasificación puede ir desde muy general como la expuesta en el ejemplo anterior o tan específica que sólo contenga estereoisómeros.

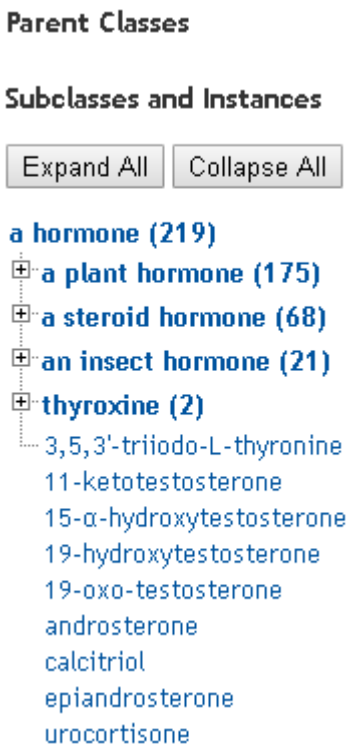


Figura 4. Ejemplo de la jerarquía de las clases en BioCyc.

La clase hormonas es una clase diversa de compuestos genéricos ya que contiene compuestos que no están familiarizados químicamente, sino funcionalmente.

Cuando se introducen compuestos genéricos cuya clasificación es ambigua, es posible que se instancien compuestos en un organismo cuando este no los produce. Para evitar esto en cada organismo se recuperaron los compuestos que tenían asignados y se validaron con los compuestos presentes en todas las reacciones del mismo organismo. La validación consistió en buscar para cada compuesto anotado en un organismo si ese mismo compuesto estaba presente en una

reacción y si era así, se anotaba como presente en el organismo, sino se ignoraba. Por ejemplo, el compuesto L-DOPA que está agrupado en la clase de los carboxilatos y alcoholes y, por lo tanto, es asignado como presente en *Escherichia coli* porque tiene una instancia de esta clase, aunque en realidad esta bacteria no sea capaz de producirla. La cantidad de clases en MetaCyc es de 343, por lo que la cantidad de compuestos que se agregan de esta forma puede ser enorme, por ejemplo, para EcoCyc son más de 1,000.

Además de los compuestos se recuperaron los identificadores taxonómicos de los organismos y sus sinónimos, para poder asignarles reacciones que no fueron propagadas a partir de MetaCyc.

3.3.5 Disponibilidad de genes

Para cada organismo se recuperaron los genes que codificaban a las enzimas que realizaban cada reacción. Al final se tomó la información de MetaCyc en la cual había información no propagada. Esta información fue curada manualmente y contiene rutas de referencia, sin embargo, en ocasiones la información de MetaCyc sobre algún organismo no está anotada en la base de datos del organismo al cual pertenece, por lo que se copiaron o propagaron las anotaciones utilizando sus clasificaciones taxonómicas como referencia. Si una reacción era realizada por enzimas compuestas de subunidades diferentes, se descargaban de BioCyc todos los genes necesarios disponibles para formar esa enzima, si estaban disponibles.

A pesar de que MetaCyc contiene información de miles de organismos y sus genes, también contiene información sobre reacciones y enzimas de organismos no presentes en la colección. Por lo tanto, las secuencias de los genes no están disponibles. En estos casos se guardó las reacciones para posteriormente descargar la información de la base de datos de KEGG, si es que existía una referencia entra ambas bases de datos. KEGG y BioCyc están enlazadas por medio de sus reacciones. En KEGG cada reacción está ligada a una enzima y cada enzima está ligada a un grupo de genes ortólogos. Que una enzima tenga más de un grupo de ortólogos en KEGG significa que esta enzima está formada por el producto de

más de un gen o que son enzimas formadas por el producto de un solo gen, que convergieron evolutivamente. Por la falta de información que provee la API de KEGG, sólo se descargaron las secuencias de los genes que poseen un solo ortólogo, esto para evitar que las enzimas encontradas fueran formadas por el producto de más de un gen, en caso contrario cada subunidad sería interpretada como una enzima completa.

Se impuso un límite de 10,000 pb para el tamaño de los genes, este límite sobrepasa por mucho el tamaño ideal para una enzima, debido a que entre mayor sea el tamaño del gen más se complica su manipulación genética (Hu *et al*., 2007), este límite se impuso por la presencia de secuencias en BioCyc que sobrepasan por mucho este tamaño, un ejemplo fue el gen GTUP-1 de *Rothia dentocariosa* ATCC 17931 que tiene 2’504,620 pb y que seguramente es un error en la anotación porque el tamaño del genoma es de 2.6 millones de pb.

3.3.6 Recuperación de los valores cinéticos para las enzimas

Para cada una de las enzimas que tenían anotado un identificador para Uniprot (72,694), se buscaron en su sitio web, la información se descargó en formato tsv y luego se procesó para hacer la separación de las Km de acuerdo con la reacción y el metabolito. Se hizo de esta manera porque a pesar de que BioCyc tiene estas anotaciones que son tomadas de Uniprot, por el momento no son incluidas en la base de datos en texto plano (Anónimo, Pathway Tools Question & Answer forum for SRI's Pathway Tools, 2017) y el procesamiento es más sencillo si se realiza desde Uniprot.

3.3.7 Coherencia entre anotaciones

La información de la base de datos está escrita en varios archivos, que se complementan entre ellos, ver **Figura 5.** Para asegurar que una reacción es realizada por un gen se busca que exista el siguiente recorrido para cada reacción:

Reacción → Reacción enzimática → Enzima → Gen

Si una reacción no cumplía este recorrido en ningún organismo y no se encuentraban los genes en KEGG, esa reacción era anotada como carente de genes, lo que significa que no se conocía ninguna enzima que la pudiera realizar, aunque la reacción se conociera.

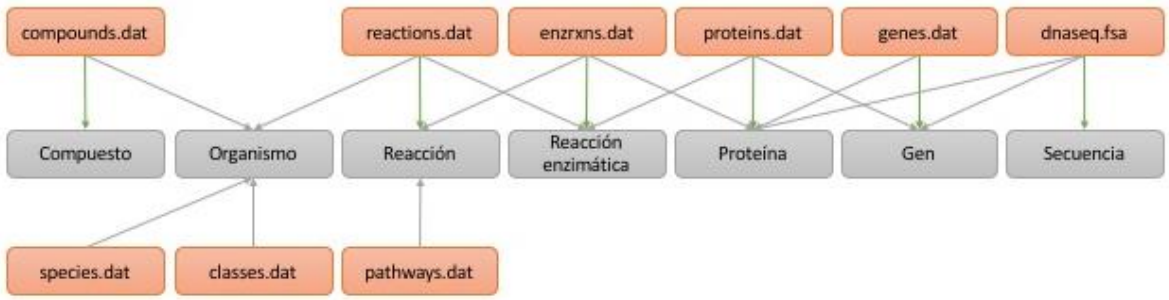


Figura 5. Recuperación de información desde los archivos de texto plano de BioCyc.

Los cuadros naranjas son archivos en texto plano de la base de datos, los cuadros grises es información extraída de la base de datos. En líneas verdes se resalta la información principal del archivo, mientras que en líneas grises esta la información enlazada al archivo.

3.4 Enumeración de rutas metabólicas

3.4.1 Retrosíntesis

La retrosíntesis normalmente se utiliza para generar rutas de síntesis para compuestos deseados, sin embargo, para este proyecto fue utilizado para reducir la cantidad de enzimas a utilizar y de esta manera reducir la demanda en el procesamiento. De forma general los pasos para llevar a cabo la retrosintesis de un compuesto son los siguientes:

1. Selección del compuesto deseado.
2. El compuesto deseado se convierte en el producto actual.
3. Se buscan todas las reacciones que producen el producto actual.
4. Las reacciones son guardadas y se generan los pares reactantes.
5. Se buscan los reactivos de estas reacciones.
6. Si se ha iterado este proceso más veces que el número máximo permitido el proceso termina aquí.

7. Cada reactivo se convierte en el producto actual si no es producido por el organismo blanco y se repite desde el paso 3.

Las reacciones y compuestos recuperados del proceso de retrosíntesis pasaron a ser el espacio metabólico extendido del organismo, a partir del cual se generaron las rutas metabólicas, teniendo la ventaja de que era menor al espacio metabólico completo de toda la base de datos y por lo tanto más rápido de analizar. Además, en lugar de utilizar las reacciones para reconstruir las rutas se utilizaron pares reactantes, en el que cada uno estaba formado por un reactivo y un producto, evitando que se generaran rutas no útiles.

3.4.2 Generación de rutas metabólicas lineales

Para reducir aún más el espacio metabólico y optimizar el tiempo de respuesta del algoritmo, la generación de rutas metabólicas se inició a partir de los compuestos presentes en el organismo, para lo cual se tomó el primer reactivo y por cada par reactante que lo contenía se generó una nueva ruta metabólica, después, los productos se volvieron reactivos y se buscaron nuevos pares reactantes, recorriendo el camino contrario a la retrosíntesis, hasta llegar al compuesto deseado.

Durante la generación de rutas metabólicas lineales se buscó que estas no contuvieran reactivos o reacciones repetidas. Además, se buscó que todas las rutas generadas empezaran con un compuesto presente en el organismo y terminaran con el compuesto deseado.

Calcular las rutas metabólicas lineales tiene inconvenientes como:

- Crecen exponencialmente con el número de pasos, pudiendo llegar a cientos de miles para un compuesto y permitiendo rutas de hasta 15 pasos.
- Las rutas pueden requerir compuestos que no se sintetizan en esa ruta y no están presentes en el organismo.
- Estas rutas pueden partir de compuestos ubicuos como pueden ser agua u oxígeno y por tanto ser falsas.

- Las rutas lineales predichas pueden en realidad ser parte de una ruta metabólica mayor.

3.4.3 Ramificación de las rutas metabólicas

Para evitar los inconvenientes mencionados de las rutas metabólicas lineales, estas se pueden ramificar. De esta forma, por cada reacción en una ruta metabólica lineal se buscaron todos sus reactivos y, si estos reactivos no se producían en esa ruta o no estaban presentes en el organismo, se le agregaba otra ruta lineal que fuera capaz de producirlo, esto se repitió hasta que todos los reactivos se puedan producir o estuvieran presentes en el organismo.

Estas rutas lineales aun así no estaban libres de errores, en ocasiones requerían de sustratos que se podían sintetizar pero que en la cadena de reacciones metabólicas se sintetizaban después de que eran requeridos. Por esta razón, primero se recrearon todos los pasos de la ruta, iniciando con los compuestos del organismo y después se agregaron las reacciones que son posibles con esos compuestos. Estos pasos se repitieron hasta que se produjera el compuesto final o no hubiera avances en las rutas; si ocurría esto último las rutas eran eliminadas. Este enfoque es el utilizado para generar los espacios metabólicos extendidos (Handorf, Ebenhöh, & Heinrich, 2005), sin embargo, aquí se utilizó para verificar cada ruta.

3.5 Calificación de las rutas metabólicas

3.5.1 Evaluación termodinámica

Debido a que la gran mayoría de las rutas no están validadas experimentalmente, se empleó la evaluación termodinámica para determinar la direccionalidad de una reacción e incluso la posibilidad de una ruta metabólica (Noor *et al*., 2014). Esto se basa principalmente en la energía libre de reacción ($\Delta_r G'$) expresada en la **Ecuación 8.**

$$\Delta_r G' = \Delta_r G'^{\circ} + RT \cdot \ln(Q) \qquad \textbf{Ecuación 8}$$

Donde:

$\Delta_r G'$ es la energía libre de reacción.

$\Delta_r G'^{\circ}$ es la energía libre estándar de la reacción.

R es la constante de los gases ideales.

T es la temperatura expresada en Kelvin.

Q es la relación de acción de masas.

Para reacciones reversibles se cumple que para:

$\Delta_r G' < 0$ la reacción se ve favorecida de izquierda a derecha.

$\Delta_r G' > 0$ la reacción se ve favorecida de derecha a izquierda.

$\Delta_r G' = 0$ la reacción está en equilibrio.

La relación de acción de masas representa la proporción que existe entre la concentración de productos y reactivos en una reacción estando la concentración dada en molaridad. Esta relación se calcula con la **Ecuación 9.**

$$Q = \frac{\prod_{i=1}^{N_P} P_i^{S_i}}{\prod_{j=1}^{N_R} R_j^{S_j}}$$ **Ecuación 9**

Donde

N_P Son los productos de la reacción

N_R Son los reactivos de la reacción

S_i Es el coeficiente estequiométrico de un producto dado

S_j Es el coeficiente estequiométrico de un reactivo dado

R_j Es un reactivo dado

P_i Es un producto dado

Para evitar confusiones en cuanto a los valores de Δ_rG' se suele utilizar con el signo contrario y se le llama Fuerza Motriz ($-\Delta_rG'$). De esta forma, entre mayor sea el valor de la Fuerza Motriz con mayor fuerza se desplazará la reacción de izquierda a derecha.

La direccionalidad prevalente de una reacción reversible se puede modificar ajustando la concentración de productos y reactivos. Si se aumenta la concentración de un compuesto se ve favorecida la dirección que lo consume. Sin embargo, dentro de los organismos no es posible variar las concentraciones libremente, de hecho, los valores regulares entre los que oscilan la mayoría de los compuestos en *Escherichia coli* es de 1 µM a 10 mM (Noor *et al*., 2014).

Aun cuando todas las reacciones de una ruta metabólica puedan suceder prevalentemente en la dirección adecuada para producir un metabolito, dentro de los límites de concentración establecidos, no significa que tal ruta metabólica sea posible ya que se debe variar la concentración de los metabolitos de forma simultánea para todas las reacciones y todas ellas deben tener una Fuerza Motriz positiva. Si no ocurre esto, la ruta metabólica no es favorable termodinámicamente (Noor *et al*., 2014).

Resolución de los estados más favorables

Para cada ruta metabólica pueden existir desde cero hasta una infinidad de concentraciones para las cuales las rutas sean posibles, por lo cual no es viable calcular estas soluciones de forma algebraica o combinatoria. Por lo tanto se recurrió a la programación lineal para resolver el estado en el cual la reacción limitante tuviera la Fuerza Motriz mayor, usando el siguiente procedimiento.

- Se creó un vector columna con los valores de las $\Delta_rG'^{\circ}$ de las reacciones de la ruta (**G°**).
- Se creó una matriz estequiométrica en la cual las columnas eran las reacciones y las filas los compuestos (**S**), para cada reacción se utilizó el coeficiente estequiométrico con signo positivo si es que se estaba generando o con signo negativo si es que era un reactivo que se estaba consumiendo.

- Se creó una matriz columna con los logaritmos naturales de las concentraciones de los compuestos (**x**).
- Se generó una nueva matriz a partir de –(**G°**+RT·$\mathbf{S^T}$·**x**).
- Se buscó mediante programación lineal un valor para el cual la reacción con una Fuerza Motriz fuera el máximo.

El problema de la programación lineal fue resuelto mediante el método Simplex, el cual es un método iterativo que recorre los vértices del poliedro multidimensional que se forma con las ecuaciones hasta que encuentra un valor máximo, si es que existía (Larson & Falvo, 2017). Para resolverlo se utilizó la librería Apache Commons Math3 de java (Apache Commons, 2016).

Cada ruta se descompuso en sus reacciones constituyentes y para cada reacción se calculó su energía libre estándar de reacción de acuerdo con la dirección en la que la reacción funcionaba para esa ruta, si la reacción funcionaba en la dirección contraria a la que fue anotada simplemente se invirtió el signo de la $\Delta_r G'°$. Con esta información y los coeficientes estequiométricos se creó la matriz de la **Ecuación 10**, dejando la columna de las concentraciones de los compuestos como incógnitas limitadas a un rango de los logaritmos de las concentraciones.

$$\Delta_r G' = \begin{bmatrix} \Delta_r G'^{\circ}{}_1 \\ \Delta_r G'^{\circ}{}_2 \\ \vdots \\ \Delta_r G'^{\circ}{}_{N_r} \end{bmatrix} + RT \cdot \begin{bmatrix} S_{1,1} & S_{2,1} \cdots S_{N_r,1} \\ S_{1,2} & S_{2,2} \cdots S_{N_r,2} \\ \vdots & \vdots \ \ddots \ \vdots \\ S_{1,N_c} S_{2,N_c} \cdots S_{N_r,N_c} \end{bmatrix}^T \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N_c} \end{bmatrix} \qquad \textbf{Ecuación 10}$$

Una vez resuelta la ecuación cada una de las filas de la matriz anterior representó una reacción en una ruta determinada, por ejemplo, para la siguiente reacción hipotética.

$$aA + \ bB \rightarrow cC$$

Se obtiene una fila en la siguiente forma:

$$\Delta_r G' = \ \Delta_r G'^{\circ} + RT(cln(x_C) - aln(x_A) - bln(x_B))$$

Comparando la **Ecuación 8** y la ecuación anterior podemos darnos cuenta de la siguiente equivalencia:

$$\ln(Q) = \ c\ln(x_C) - a\ln(x_A) - b\ln(x_B)$$

Esta igualdad se debe a las propiedades de los logaritmos. Esta equivalencia es importante debido a que la optimización lineal sólo es aplicable a ecuaciones lineales con la forma de la **Ecuación 7** y de esta forma la ecuación cumple con este requisito y puede generar programas lineales.

Para cada una de las reacciones de una ruta se creó un programa lineal en el cual se buscaba hacer la optimización para encontrar el valor máximo de la Fuerza Motriz tal que se cumplieran las siguientes condiciones:

-El valor para cada una de las concentraciones de los compuestos (x_i) se limitó como:

$$\ln(0.000001\ M) \leq \ x_i \leq \ln(0.01\ M)$$

-El valor de $-\Delta_r G'$ para la reacción que se estaba probando tenía que ser menor o igual al valor de $-\Delta_r G'_i$ de cada una de las reacciones de la ruta.

$$-\Delta_r G' \leq \ -\Delta_r G'_i$$

De los programas lineales que tuvieron solución se encuentró cual es el que tiene el valor mayor y este se estableció como la máxima fuerza motriz para la reacción limitante de la ruta, el valor calculado esta expresado en Jules.

Significado de la evaluación termodinámica

El resultado que produjo el programa lineal representa el $-\Delta_r G'$ de la reacción limitante para las mejores concentraciones en la ruta metabólica, por lo que la ruta metabólica con un mayor resultado requerirá de una menor concentración de enzimas y por tanto un menor desgaste por parte del organismo, suponiendo que todas las enzimas utilizadas tengan la misma eficiencia catalítica.

En caso de que no se haya encontrado ninguna solución para la ruta significa que esa ruta no es posible termodinámicamente. Sin embargo, se debe tomar en cuenta que los valores termodinámicos no fueron obtenidos de forma experimental, sino que muchos de ellos se calcularon mediante el método de contribución de los componentes, en el cual se descomponen a las moléculas en grupos de átomos y luego se calcula la energía libre de formación a partir de estos valores para después calcular la energía libre de reacción, lo que puede introducir errores en los análisis. Por esta razón, las rutas no posibles termodinámicamente se mantienen a reserva de que se demuestren experimentalmente que no son posibles, además de que para algunos compuestos o reactivos no es posible, por ahora, asignarles valores termodinámicos debido a su complejidad.

3.5.2 Similitud del compuesto precursor con el compuesto deseado

McShan y colaboradores en 2003 propusieron un algoritmo para la búsqueda heurística de rutas metabólicas, este algoritmo define a los compuestos como vectores de átomos y enlaces, mientras que a las reacciones químicas como transformaciones que cambian la cantidad de átomos y enlaces entre un compuesto y su precursor. Para cada reacción se le asigna un costo de transformación química el cual representa la cantidad de átomos y enlaces que gana o pierde un compuesto durante una reacción.

Para determinar cuáles rutas eran más prometedoras para producir un compuesto ellos evaluaron si una reacción agregada a una ruta en construcción disminuía el costo de las transformaciones químicas entre dos compuestos o lo aumentaba, para esto ellos utilizaron la **Ecuación 11** y buscaron que el resultado siempre sea el menor.

$$F(0,m,L) = \sum_{i=1}^{i=m} \left(\left|x^i - x^{i-1}\right|\right) + \left|x^m - x^L\right| \qquad \textbf{Ecuación 11}$$

Donde $F(0,m,L)$ es el costo para producir un compuesto utilizando el intermediario x^m a partir del compuesto inicial x^1, $\left|x^i - x^{i-1}\right|$ es la distancia Manhattan del

compesto inicial al intermediario x^m, y $|x^m - x^L|$ es la distancia Manhattan del intermediario al compuesto deseado x^L.

La distancia Manhattan es llamada así por la ciudad homónima debido a su forma cuadriculada, esta distancia también se le conoce como Taxicab porque es utilizada para problemas en los cuales se requiere reducir la distancia recorrida por un taxista que lleva un pasajero de un punto a otro de la ciudad, en la **Figura 6** se muestra una representación gráfica de la **Ecuación 11** con ejemplos.

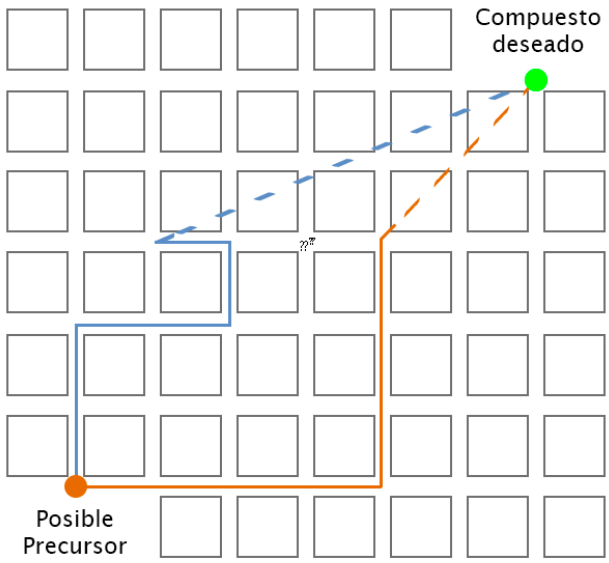


Figura 6. Representación gráfica del costo de transformación química.

En esta figura se muestra la representación de dos rutas en proceso de construcción, ambas rutas llevan el mismo costo de transformación química (Líneas continuas), sin embargo, la ruta naranja está más cerca del compuesto deseado que la ruta azul (Líneas discontinuas). En el algoritmo de McShan se le da preferencia a la ruta naranja. Las líneas continuas representan el primer término de la **Ecuación 11**, mientras que las líneas discontinuas representan el segundo término. Cada arista de un cuadrado representa a una reacción, y la longitud de la arista representa el costo de la transformación química.

Si bien el costo real de una ruta metabólica está definido por un gran número de factores, actualmente no se tiene suficiente información para evaluarlas tomándolos todos en cuenta, por lo que McShan y colaboradores sugieren que utilizar el costo

en las transformaciones químicas es congruente con la tendencia en los organismos en optimizar su crecimiento (McShan *et al*., 2003).

Para determinar la similitud que tiene un compuesto deseado con el compuesto precursor presente en el organismo se calculó el costo de las transformaciones químicas. Para, esto se utilizó la distancia Manhattan, en la cual se contó el número de átomos que se incorporaron o salieron de un compuesto durante todas las reacciones. Este cálculo permitió tanto penalizar rutas con pasos redundantes o con pasos innecesarios, como buscar rutas con un menor número de reacciones ya que entre más diferente sea un compuesto, una mayor cantidad de modificaciones se necesitan para producir el compuesto final (McShan, Rao, & Shah, 2003).

Para cada una de las rutas se tomaron los pares reactantes generados por el programa y a partir de ellos se obtuvieron sus formulas químicas. Para cada uno de los átomos se obtuvo la diferencia y se sumó, el resultado final fue el número de átomos que ganó o perdió la molécula. Esta distancia se calcula con la **Ecuación 12**, esta ecuación es una adaptación de la ecuación para calcular transformaciones químicas utilizada por McShan *y* colaboradores en 2003, aquí no se incluyen los enlaces químico debido a que no es sencillo recuperarlos de la colección BioCyc, además de que no interfieren en la utilización de biomasa. La **Ecuación 12** se deriva de la **Ecuación 11**, cuando el intermediario es igual al compuesto final.

$$CTQ = \sum_{i=1}^{N_r} \sum_{j=1}^{N_A} |AP_{ij} - AR_{ij}| \qquad \textbf{Ecuación 12}$$

Donde CTQ es el costo de las transformaciones químicas, N_r son las reacciones de la ruta, N_A son los átomos de la ruta, AP es el número de átomos en el producto actual y AR es el número de átomos en el reactivo actual.

3.6 Búsqueda de enzimas y genes

Para reducir la carga representada por las enzimas que necesitan ser expresadas heterologamente en el organismo huésped para producir un compuesto, se

ordenaron primero las que eran capaces de realizar las reacciones compuestas con el mayor número de sub-reacciones, evitando de esta manera tener redundancia en reacciones o un mayor número de genes y reacciones involucradas.

Las enzimas encontradas se ordenaron con un algoritmo de ordenación de burbuja (Bubble Sorting) que utiliza la siguiente jerarquía:

- Enzimas con reacciones curadas manualmente.
    - Enzimas con un valor menor Km.
        - Menor número de genes requeridos por la enzima.
            - Número de reacciones realizadas en la ruta.
                - Tamaño de los genes requeridos para la enzima.

3.7 Estimación de los espacios metabólicos extendidos para cada organismo

En Java, se escribió un programa que extendió el espacio metabólico de cada organismo, utilizando las reacciones y compuestos de BioCyc. Para esto el programa leyó la base de datos creada localmente en este trabajo y para cada organismo recuperó los compuestos que es capaz de producir. A partir de ellos buscó para qué reacciones existían todos los precursores en el organismo y creó el primer espacio metabólico extendido en el cual estaban presentes las reacciones que producía el organismo y los productos de las reacciones que tenían disponibles todos sus precursores, este proceso se repitió utilizando el espacio metabólico extendido como fuente de precursores en lugar de los compuestos del organismo. El proceso se detuvo hasta que se dejaban de agregar nuevos compuestos al espacio metabólico extendido, y este se consideró el espacio final. El número de espacios metabólicos indica cuál es la ruta lineal más larga para la producción de un metabolito, además, la cantidad de compuestos finales es el potencial metabólico que tiene ese organismo para producir metabolitos de forma heteróloga. El proceso anteriormente descrito fue utilizado por Handorf y colaboradores en 2005 para la generación de espacios metabólicos extendidos.

3.8 Análisis de la red metabólica de BioCyc

Para determinar cuáles son los principales compuestos en la red metabólica creada a partir de las reacciones de BioCyc, se creó un grafo múltiple dirigido, de acuerdo a las siguientes consideraciones:

Se revisó manualmente cada reacción para crear una lista de compuestos altamente conectados o que fueran pequeñas moléculas ubicuas, como ATP, H2O y CO2 (para ver la lista completa consultar el **Apéndice 1**). Para saber si estas moléculas no tenían un papel principal en la reacción se revisaron los mapas atómicos de las reacciones mostrados en el sitio web de BioCyc, donde se puede ver el destino de cada átomo en una reacción, de esta forma se pudo saber si una molécula funcionaba como transportadora de grupos funcionales; se encontraron los metabolitos con un índice de carga mayor que fueran moléculas pequeñas como sales; se encontraron pares reactantes que estaban presentes en varias reacciones lo que indicaría que funcionaban como transportadores de grupos o energía; parte de la clasificación fue arbitraría porque no existe una definición formal de estos compuestos. Estos compuestos le restan significado biológico a la red debido a que están tan altamente conectados que los algoritmos para procesar redes saltan por ellos y encuentran rutas más cortas que en realidad no tienen ningún sentido biológico (Ma & Zeng, 2003).

Para cada una de las reacciones se crearon pares reactantes utilizando la combinación de cada compuesto con cada producto, siempre y cuando el compuesto no estuviera en la lista de compuestos del **Apéndice 1** o la reacción tuviera una anotación manual.

Para cada una de las reacciones se revisaron los pares reactantes que se producirían por el paso anterior y en caso de producir pares reactantes equivocados se procedía a hacer la anotación manual, por ejemplo, cuando un reactivo sólo contribuía a formar algunos productos y no a todos (Junker & Schreiber, 2008). A continuación estos casos se detallan con ejemplos resueltos en la **Tabla 5** y **Tabla 6**.

Tabla 5. Creación de pares reactantes a partir de una reacción donde existe un transportador de grupos funcionales.

Reacción RXN-8149 de MetaCyc, ver Figura 7.			
Tebaína + 2-oxoglutarato + O2 → Oripavina + Formaldehído + Succinato + CO2			
Pares reactantes	Tebaína	→	Oripavina
	Tebaína	→	Formaldehído
	2-oxoglutarato	→	Succinato
Las moléculas de Dióxido de Carbono y Oxigeno se ignoran porque son moléculas pequeñas y ubicuas que restarían significado biológico a la red. Si se tomara en cuenta aquí la molécula de oxígeno al momento de procesar las redes el algoritmo predeciría que a partir del oxígeno se puede producir oripavina que, aunque contribuye a su formación no es el sustrato principal.			

Tabla 6. Creación de pares reactantes a partir de reacciones involucradas en el metabolismo de compuestos transportadores de grupos funcionales.

Reacción BTUR2-RXN de MetaCyc, ver Figura *8*.			
Cobinamida + ATP → Adenosilcobinamida + PPP			
Pares reactantes	Cobinamida	→	Adenosilcobinamida
	ATP	→	Adenosilcobinamida
En esta reacción el ATP que es un metabolito altamente conectado no participa en la transferencia de energía o grupos funcionales, sino que sirve de precursor en la síntesis de adenosilcobinamida, durante la ruta de recuperación de la adenosina.			

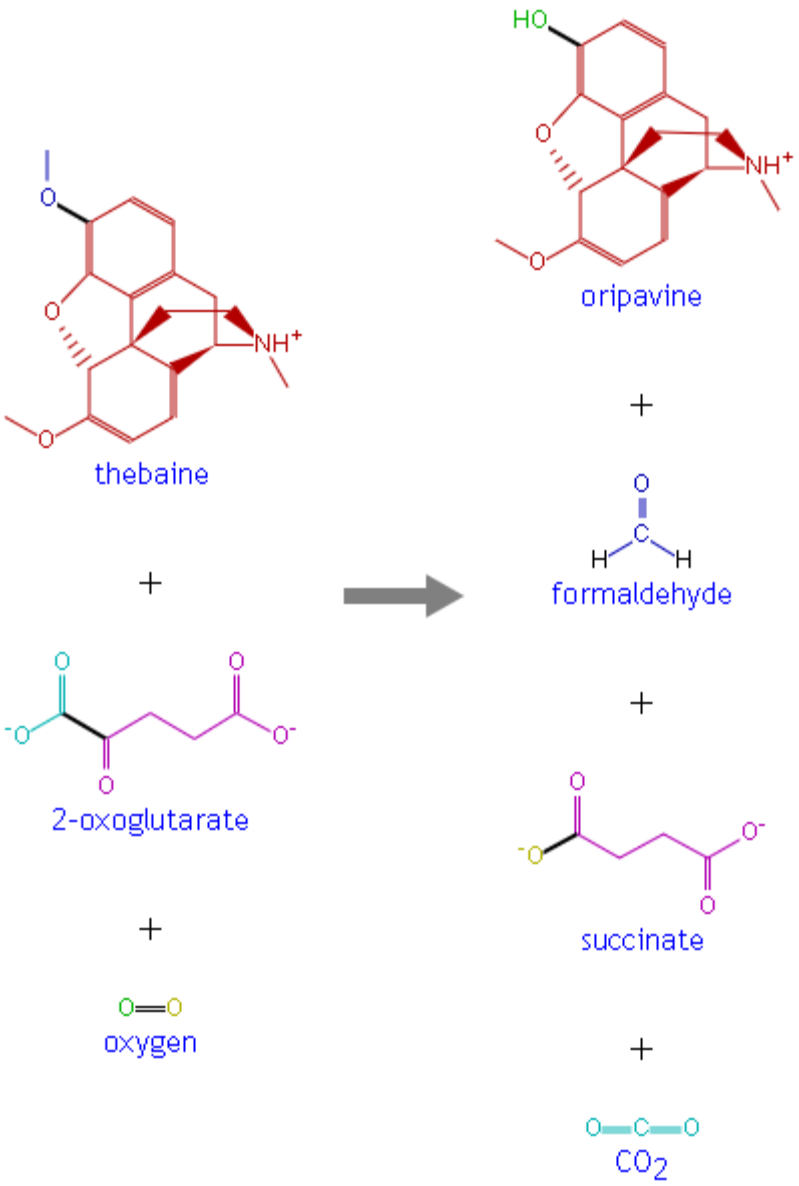


Figura 7. Reacción RXN-8149 de MetaCyc (Caspi *et al*., 2016).

El color de los átomos indica a que producto contribuyeron en la formación, aquí se observa que el 2-oxoglutarato produjo el succinato y dióxido de carbono, mientras que la tebeina se dividió en formaldehído y oripavina. Ignorando las moléculas pequeñas (oxígeno y dióxido de carbono) se pueden generar tres pares reactantes de esta reacción.

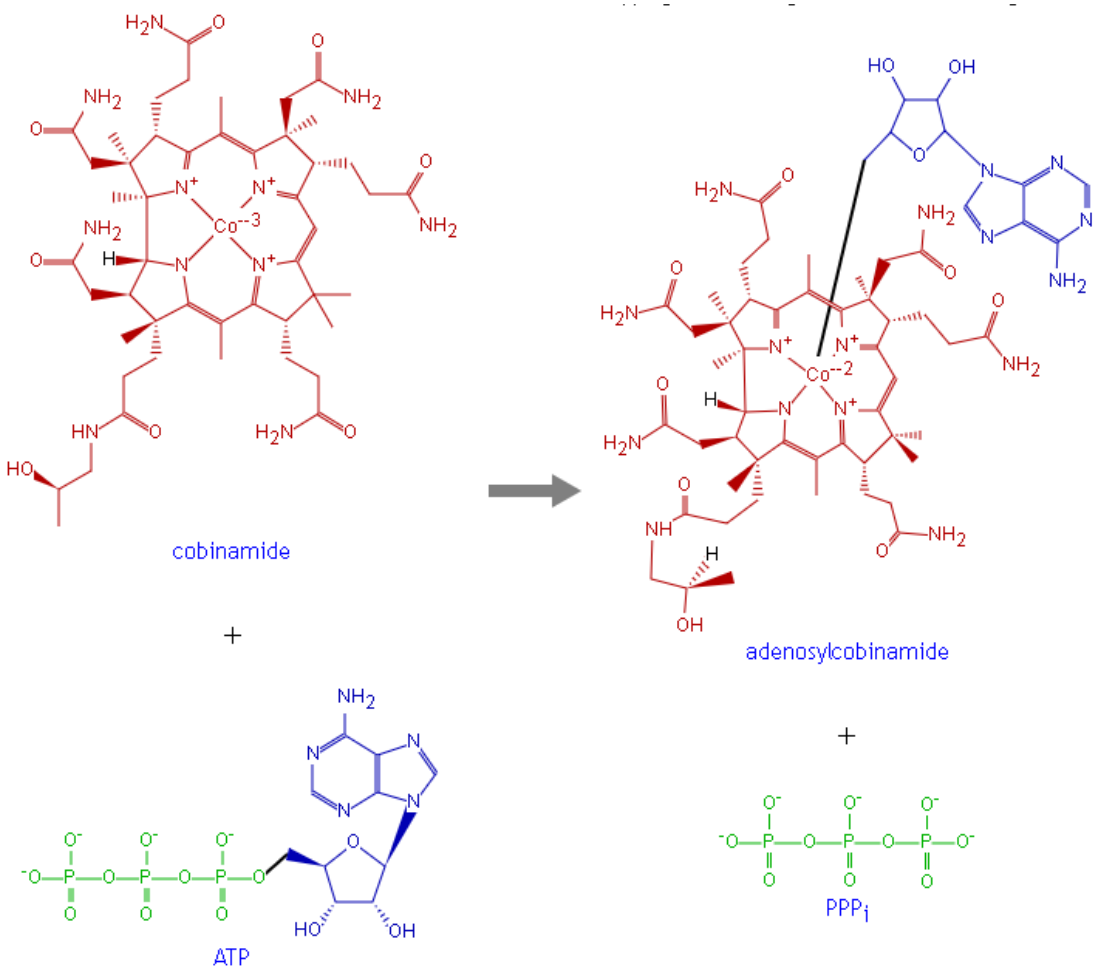


Figura 8. Reacción BTUR2-RXN de MetaCyc con estructuras químicas (Caspi *et al*., 2016).

En esta reacción se observa que el ATP no participa como un transportador de energía sino como un precursor de la adenosilcobanamida. A partir de esta reacción se pueden generar dos pares reactantes ya que tiene dos precursores principales y un producto, el PPPi se descarta por ser un compuesto pequeño y ubicuo.

Con estos pares reactantes se creó un grafo dirigido en el cual se indicó la dirección de la reacción mediante aristas y cada compuesto fue representado por un nodo. Para anotar si una reacción es reversible se agregaron dos aristas en sentido inverso entre el reactivo y el producto, esto convirtió al grafo dirigido en un multígrafo dirigido porque tiene pares de nodos conectados por más de una arista. Además, los compuestos que no estaban conectados a otros compuestos no fueron agregados a la red (ver **Figura 17)**.

A partir del archivo “id2rn” de la carpeta llamada “redes” que contiene las reacciones de BioCyc y utilizando el programa CHetCount de la carpeta “proyectos”, que también simula las redes expandidas (ver **Apéndice 4**), se crean los archivos con las redes metabólicas en dos formatos (network.graphml y pair.dat), el primero siguiendo las especificaciones del formato GraphML (GraphML team, 2016), para visualizarse en Cytoscape y el segundo es una lista de comandos en texto plano para Python que utiliza el paquete NetworkX para crear una red en una variable llamada “G”. CHetCount requiere de la carpeta que generó el programa DBBuilder para poder funcionar.

Importando el archivo “network.graphml” a Cytoscape se utilizó la herramienta “Analyze Network” para redes dirigidas, consiguiendo de esta forma el diámetro de la red, el número de componentes conectados y la distribución de las rutas más cortas.

Ejecutando los comandos del archivo “pair.dat” en una consola de Python se creó un multígrafo dirigido con en una variable llamada “G” a partir de la cual se calcularon los índices de centralidad de la carga y de cercanía para los compuestos de acuerdo con el manual de NetworkX (Hagberg, Schult, & Swart, 2008). Algo importante a resaltar es que NetworkX no permite calcular parámetros como el diámetro de la red para redes con varios componentes.

4 Resultados

Se logró construir un programa para la reconstrucción de rutas metabólicas partiendo de las reacciones presentes en la colección BioCyc (Ver **Apéndice 5**). También se logró implementar un algoritmo para la calificación de cada una de las rutas metabólicas de acuerdo con sus propiedades termodinámicas y a la conservación de los átomos del compuesto original. Además de usar principalmente información de BioCyc, también se complementaron los genes faltantes con información de KEGG cuando fue posible y se combinó con información de Uniprot para hacerla más confiable. Con este programa se pudo aumentar la cantidad de metabolitos que produce *E. coli* de 1,093 a 4,361 y se pudo conseguir resultados similares para otros miles de organismos, permitiendo resaltar otros organismos que se puedan usar como chasis para producir compuestos de alto valor.

Las limitantes que encontramos para que todos los organismos pudieran producir todos los compuestos presentes en la base de datos local fueron:

1. La presencia de más de un componente conectado en la red de reacciones de BioCyc, en la **Figura 9** se puede observar como existen compuestos que formaron grupos aislados del componente de la red que contiene a la mayoría de los compuestos, como estos compuestos están aislado significa que no existen enzimas que puedan producirlos.
2. Como la red es dirigida, el hecho de que un compuesto “A” pueda producir un compuesto “B” no significa que el compuesto “B” pueda producir al compuesto “A”. Esto es importante porque BioCyc contiene rutas de degradación, por lo que estas rutas indican la transformación de un compuesto exógeno del organismo hasta un compuesto que produce el organismo y por esta razón no se pueden producir para todos los organismos.
3. Cada organismo produce una cantidad diferente de compuestos por lo que la cantidad de rutas que puede conectar son diferentes.

Para aumentar la cantidad de compuestos que pueden producir todos los organismos es necesario agregar manualmente las reacciones que hagan falta para que la red este formada por un componente único conectado. Si bien es poco

probable que se puedan producir todos los compuestos en todos los organismos debido a la existencia de reacciones irreversibles o rutas alternas que las complementen, un hecho importante aquí es que se podrían aprovechar las rutas de degradación reportadas por BioCyc para introducirlas en organismos y estos puedan ser utilizados en biorremediación o utilizar fuentes alternativas de carbono.

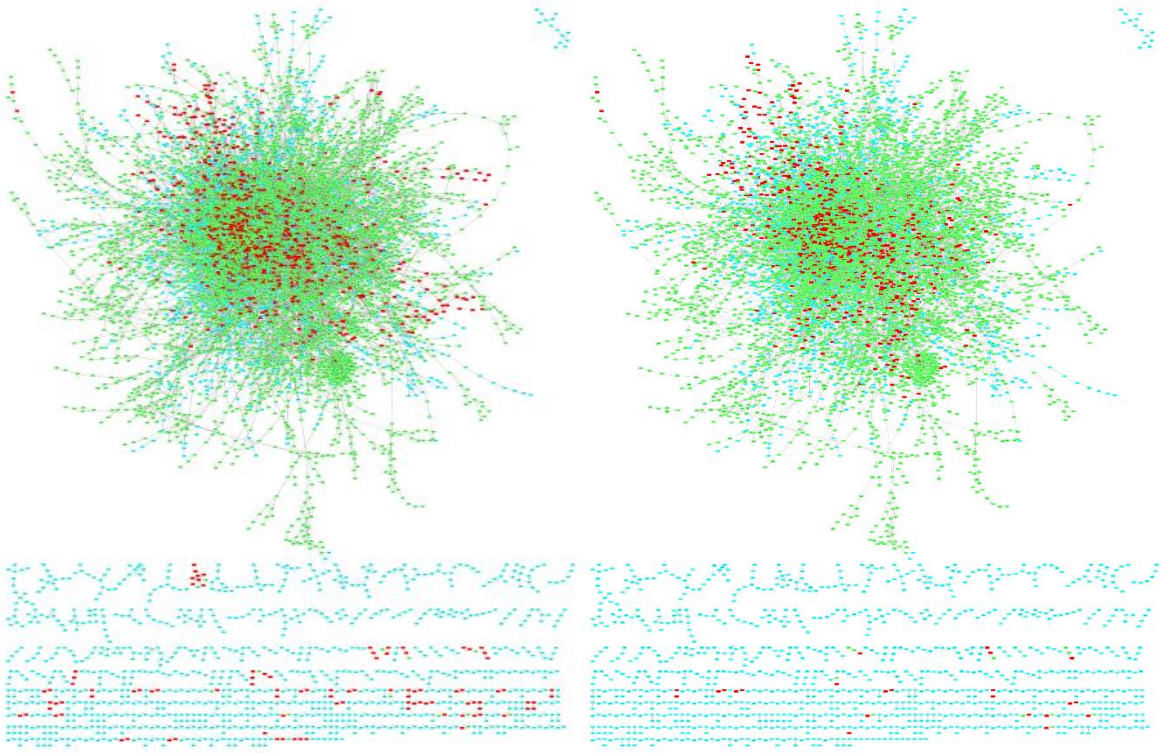


Figura 9. Comparación entre la red metabólica extendida para dos organismos hipotéticos.

En color rojo se indican los compuestos que pueden ser producidos por el organismo hipotético, en verde los compuestos que puede producir agregando enzimas hetorólogas, y en azul los compuestos que no puede producir.

La base de datos local generada a partir de BioCyc comprende a 7,591 organismos, con un total de 12,191 reacciones, 8,722 compuestos y 3’414,512 genes. A los genes de BioCyc se agregaron 38,888 genes de la base de datos KEGG GENES, para complementar la información de 319 enzimas que no tienen un gen anotado en BioCyc, con esto serían 5,777 las reacciones relacionadas a un gen, quedando sólo 6,414 reacciones asociadas a enzimas, pero sin un gen anotado en la base de

datos. De los 3’414,512 genes de BioCyc sólo se encontró registro para 72,694 genes en la base de datos Uniprot y de esta base de datos se obtuvieron 1,444 valores de Km para distintos sustratos y enzimas. El número de reacciones anotadas que pueden ocurrir de forma espontánea fueron 441.

Algunas reacciones no contienen un gen asociado, pero se mantuvieron en la base de datos y fueron utilizadas para la reconstrucción de rutas metabólicas, porque BioCyc a pesar de tener muchos años de trabajo invertidos no está completa y aún es necesario seguir agregando información. Si el usuario está interesado en una ruta en particular que utiliza una reacción sin un gen anotado podría buscarla en otras bases de datos o en artículos.

Tabla 7. Resumen de las fuentes de la base de datos local.

Información	BioCyc	KEGG	Uniprot
Compuestos	8,722	-	-
Reacciones*	12,191	219	-
Genes	3’414,512	38,888	-
Organismos*	7,591	4,790	-
Uniprot ID	72,694	-	-
Km	-	-	1,444

*Puede haber una intersección entre las bases de datos

4.1 Reconstrucción de rutas metabólicas

Para comprobar la capacidad del programa en la reconstrucción de rutas metabólicas se recuperaron todos los compuestos presentes en las rutas de referencia de MetaCyc correspondientes a la clase *“Secondary Metabolites Biosynthesis”*. De estos compuestos se eliminaron los presentes en *Escherichia coli*, quedando 240 compuestos. El programa pudo reconstruir al menos una ruta para 185 de ellos (77%) mientras que, el resto de las rutas no pudieron ser predichas debido al tipo de anotación existente en BioCyc para enzimas promiscuas, las cuales no se tomaron en cuenta debido a la carencia de información más específica sobre esas reacciones, ver **Apéndice 3**. Para cada una de las rutas predichas para los 185 compuestos, al menos una de ellas coincidió con las rutas de referencia.

4.1.1 Producción de retinal en *Escherichia coli*

Los retinoides son un grupo de moléculas lipofílicas utilizadas en cosméticos y tratamientos para enfermedades de la piel. Estos compuestos son producidos a partir del retinal que es derivado a su vez de la ruta de los carotenoides (Jang *et al*., 2011). Para simular la ruta de síntesis del retinal se seleccionó a *Escherichia coli* MG1655 como chasis, permitiéndole expandir su red metabólica hasta en 12 reacciones. Se obtuvieron 84 posibles rutas, de las cuales a pesar de todas tener valores termodinámicos favorables, 74 contenían una enzima para la cual no se conoce un gen en la base de datos local.

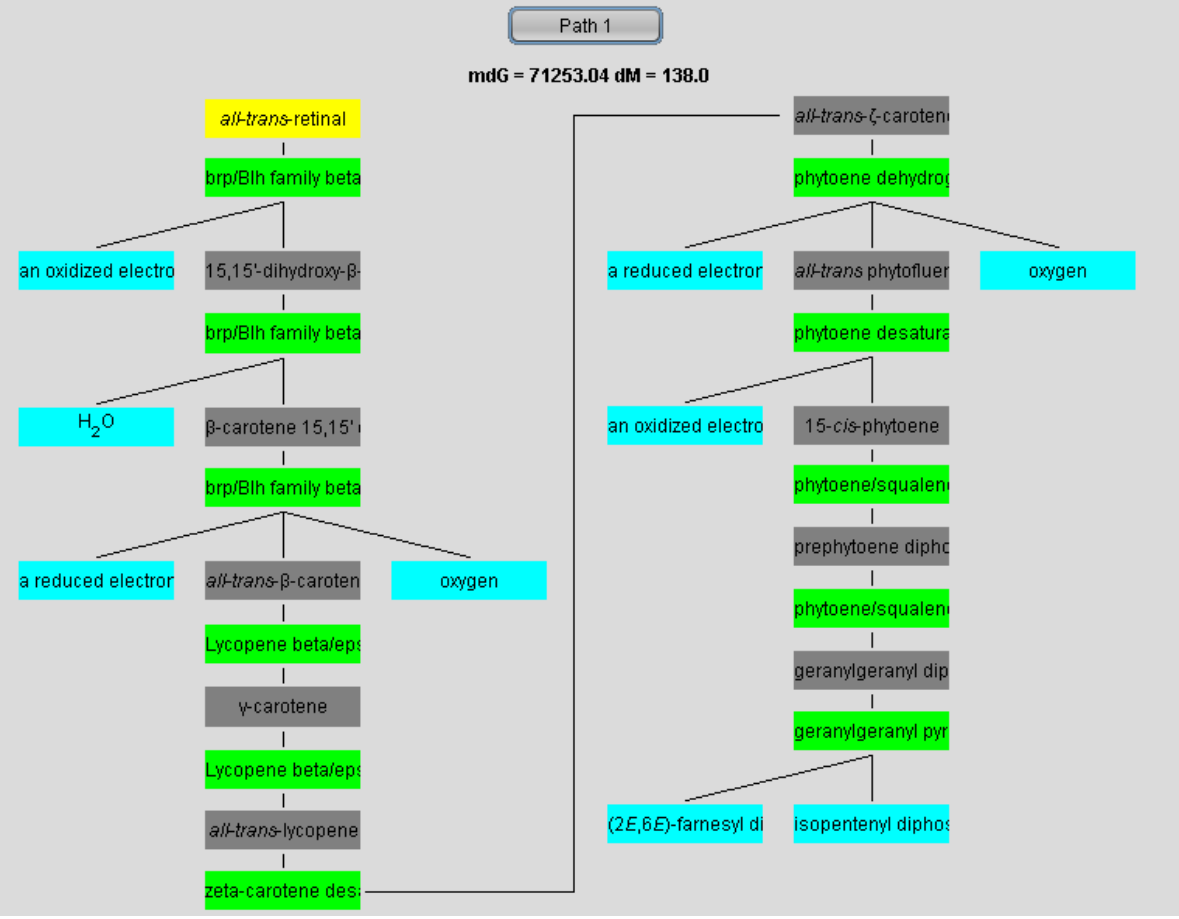


Figura 10. Ruta para la producción de retinal en *E. coli*.
Los rectángulos verdes representan las enzimas, los azules son los compuestos producidos por *E. coli*, en gris los compuestos que no puede producir *E. coli* y en amarillo el compuesto deseado.

En la **Figura 10** se muestra la ruta evaluada como la más viable de acuerdo con los valores termodinámicos, coincide perfectamente con la ruta utilizada por Jang y colaboradores en 2011. Las otras rutas generadas difieren en la forma en que se sintetiza el geranilgeranil difosfato, por ejemplo, es posible sintetizarlo utilizando el compuesto heptaprenil difosfato como intermediario y, a pesar de que tienen la misma posibilidad termodinámica de realizarse, en esta ruta (**Figura 11**) existe un paso extra y por esa razón la cantidad de átomos que entran y salen de la ruta es mayor. Además de estas variaciones en las rutas se encuentran rutas repetidas en las cuales la cantidad de enzimas necesarias varía, ya que algunas de las enzimas pueden llevar a cabo más de una reacción. Esta anotación se toma en cuenta por el algoritmo que es capaz de reconocer las reacciones compuestas de BioCyc y sus constituyentes para buscar cuáles son iguales, sin embargo, no todas las bases de datos tienen este tipo de anotaciones y por esta razón existe redundancia en las rutas generadas (SRI International, 2016).

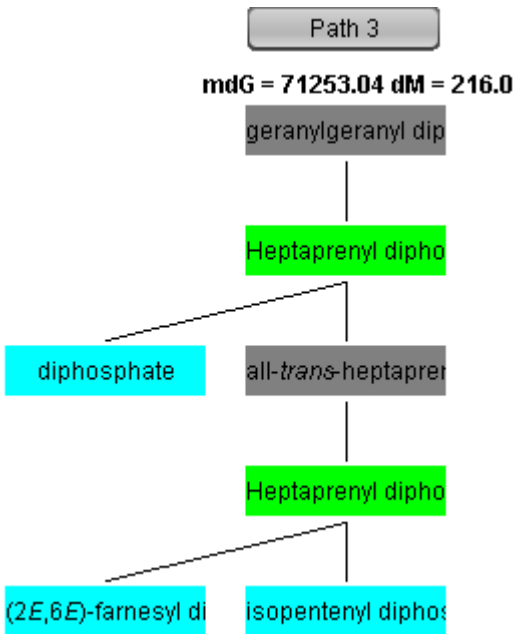


Figura 11. Ruta alternativa generada con nuestro algoritmo para la producción de geranilgeranil difosfato en la producción de retinal (ruta recortada).

Los rectángulos verdes representan las enzimas de las que carece *E. coli*, los azules a los compuestos que puede producir y en gris los compuestos que no puede producir.

4.1.2 Producción de 1,3-diaminopropano en *Escherichia coli*

La diamina de tres carbonos llamada 1,3-diaminopropano (1,3-DAP) es utilizada como monómero en la fabricación de plásticos. Naturalmente *Pseudomonas aeruginosa* (**Figura 12**) y *Acinebacter baumannii* (**Figura 13**) son capaces de producirlas en bajas concentraciones, sin embargo, ambas son patógenas y por esta razón no es viable utilizarlas para la producción industrial. Por esta razón Chae, *et al*. en 2015 llevaron a cabo la producción de 1,3-DAP en *E. coli* utilizando las rutas de *P. aeruginosa* y *A. baumannii*.

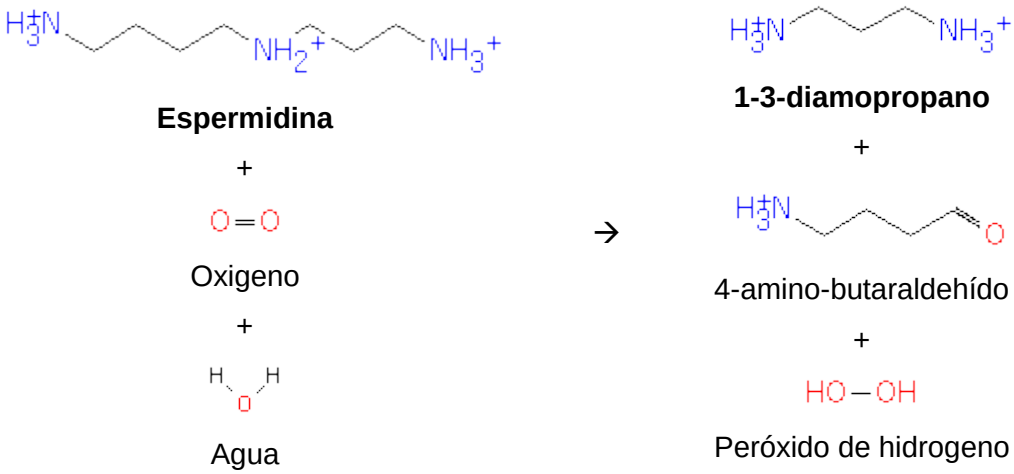


Figura 12. Síntesis de 1,3-DAP en *P. aeruginosa (*Chae *et al*., 2015).

En negritas están los compuestos principales de la ruta, aquí el 4-amino-butaraldehído se considera como un compuesto saliente.

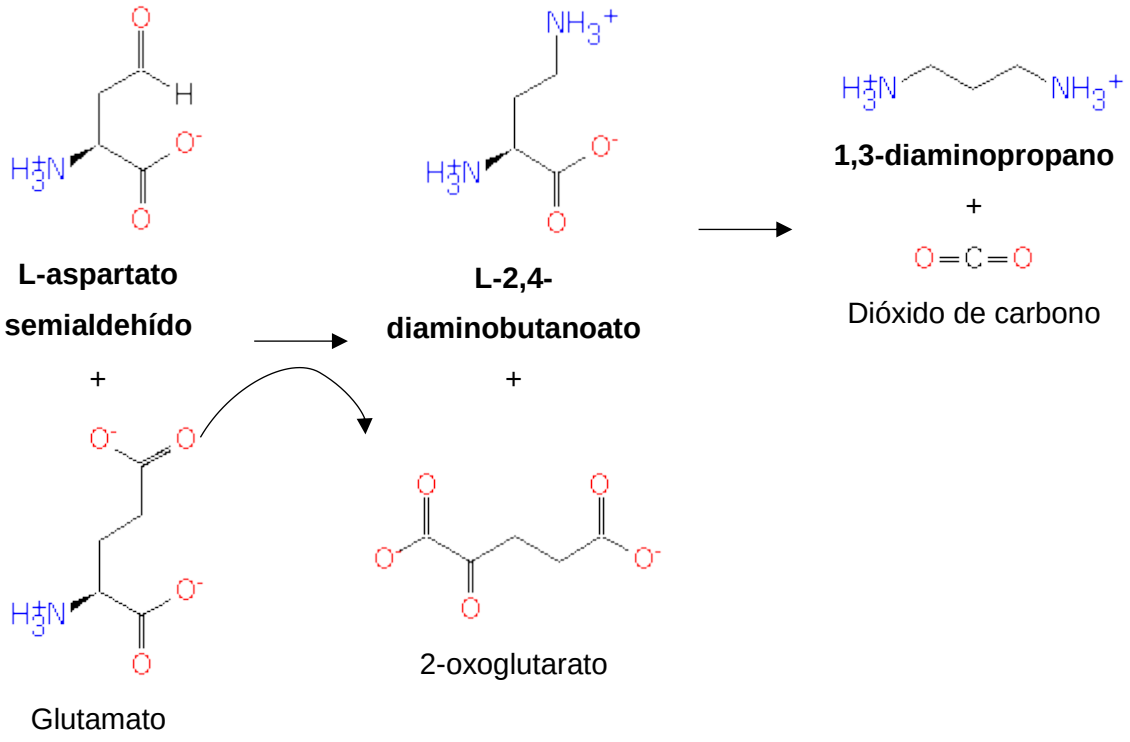


Figura 13. Síntesis de 1,3-DAP en *A. baumannii* (Chae *et al.*, 2015)

En negritas están los compuestos principales de la ruta. En esta reacción el glutamato actúa como un donador de grupos amino por esta razón no es principal.

Cuando se reconstruyeron estas rutas utilizando a *E. coli* como organismo chasis y permitiendo rutas con un máximo de 4 reacciones, obtenemos 11 rutas de las cuales una de ellas no tiene un gen asociado a una de sus reacciones. Las rutas utilizadas en el artículo son predichas por Chae *et al.* en la posición 1 y 3, la posición 2 es ocupada por una variante de la ruta 1. De las rutas creadas, 5 de ellas son termodinámicamente favorables (ver **Figura 14**) y de estas rutas se pueden considerar que en realidad son dos ya que en ellas hay enzimas que realizan la misma reacción, pero con distintas moléculas transportadoras de electrones y grupos amino.

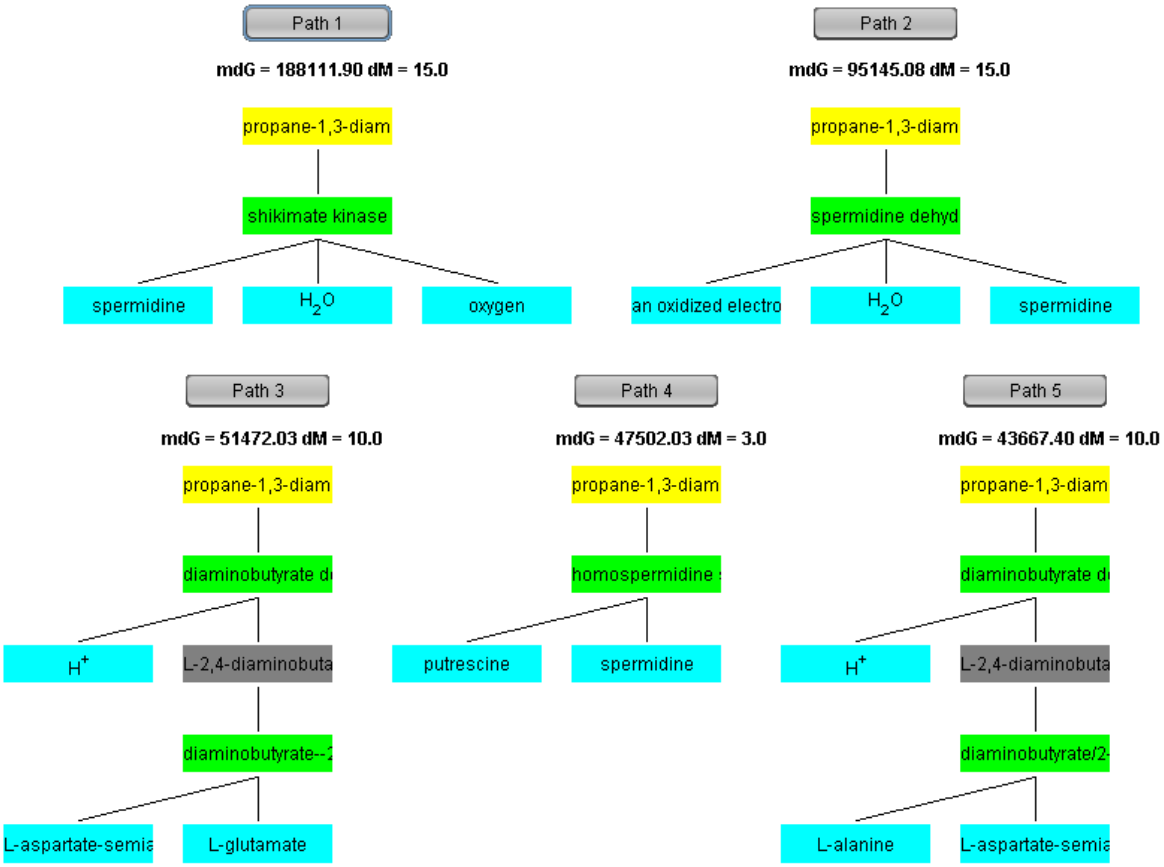


Figura *14*. Rutas termodinámicamente favorables para la producción de 1,3-DAP en E. coli.

Estas rutas son termodinámicamente favorables para la producción de 1,3-DAP, se omiten las rutas que no son favorables termodinámicamente o carecen de genes. Como se puede observar las rutas 1, 2 y 4 se producen a partir de espermidina, mientras que las rutas 3 y 5 se producen a partir de L-aspartato-semialdehído, la diferencia entre estos dos grupos de rutas es el uso de moléculas transportadoras de grupos funcionales. Ambos grupos rutas fueron probadas por Chae y colaboradores en 2015.

4.1.3 Producción de p-hidroxibenzoato en *Pseudomonas putida*

El p-hidroxibenzoato es un químico utilizado en la producción de cristal líquido, y sus derivados alquilados son utilizados como preservantes de cosméticos (Verhoef, Ruijssenaars, de Bont, & Wery, 2007). Actualmente este compuesto es producido a partir de fenol a través de la reacción de Kolbe-Schmit la cual requiere de temperatura y presión extremas. Se ha encontrado que para *E. coli* el p-

hidroxibenzoato es tóxico, por esta razón se ha tratado de producir en *Pseudomonas putida* que tiene una mayor tolerancia al producto (Verhoef *et al*., 2017).

Utilizando el programa desarrollado aquí se encuentra que el p-hidroxibenzoato es un compuesto producido por *Pseudomonas putida* de forma nativa, a pesar de que Verhoef y colaboradores requieren introducir la enzima fenilalanina amonio liasa (Pal) proveniente de la levadura *Rhodosporium toruloides* para convertir el aminoácido L-tirosina en p-coumarato; lo plantearon así porque *Pseudomonas putida* posteriormente es capaz de producir p-hidroxibenzoato a partir de p-coumarato utilizando una enzima nativa (ver **Figura 15**), y nuestro algoritmo no toma en cuenta las reacciones internas del organismo a utilizar, debido a que no es posible determinar a partir de qué metabolito es adecuado iniciar la producción. En este ejemplo parece muy claro que el algoritmo falla porque es necesario introducir un gen más para producir p-coumarato que es el precursor de p-hidroxibenzoato, o es necesario agregar p-coumarato al medio para que la bacteria lo transforme a p-hidroxibenzoato. Las definiciones que existen para considerar que un compuesto está presente en un organismo no son compatibles con los requerimientos de nuestro algoritmo. Algunas bases de datos como BioCyc y KEGG anotan compuestos en los organismos siempre y cuando estos compuestos participen en alguna reacción realizada por una enzima presente en el organismo, sin importar que los compuestos estén siendo degradados en lugar de sintetizados. Agregar compuestos que se degradan puede permitirle a la base de datos considerar más compuestos de rutas que no están completamente anotadas en el organismo. Sin embargo, un problema recurrente que hemos encontrado es la presencia de enzimas que naturalmente no realizan una reacción, pero que hay evidencia experimental de que se han llevado a cabo *in vitro* por lo que los compuestos de estas reacciones son anotados en la base de datos, a pesar de que esta evidencia indique que se desconoce la función en el organismo. Un ejemplo de lo anteriormente expuesto es un caso anotado en la base de datos ECMDB que es una base de datos que estudia el metaboloma de *E. coli* MG1655 (Guo *et al*., 2013), donde las curaciones son realizadas manuales y tienen estándares muy altos de

calidad para verificarlas. En esta base de datos se considera la L-DOPA como un compuesto presente en *E. coli*. Sin embargo, si revisamos las referencias sobre esta anotación, incluso en los mismos comentarios de la base de datos, esta reacción no ocurre naturalmente en *E. coli*. Podemos darnos cuenta de que esta anotación se considera por la presencia del gen *ygiD* que codifica para una 4,5-DOPA estradiol dioxigenasa (con el registro P24197 en Uniprot), en el artículo original Gandía-Herrero y García-Carmona en 2014 indican que esta reacción se llevó a cabo *in vitro* y que se desconoce que reacción realiza en *E. coli*. Estas anotaciones provocan que nuestro algoritmo prediga rutas más cortas a partir de compuestos que un organismo no produce naturalmente o que no están conectados al metabolismo central.

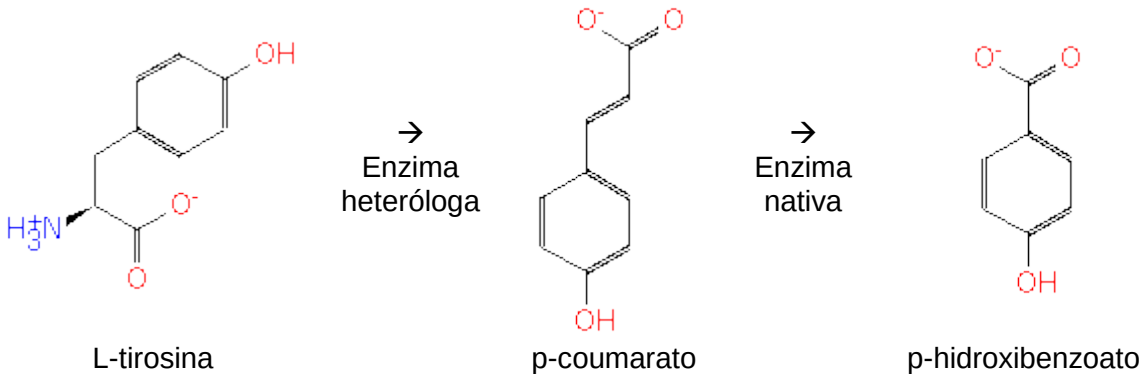


Figura 15. Producción de p-hidroxibenzoato en *Pseudomonas putida*.

Pseudomonas putida puede producir p-hidroxibenzoato a partir de p-coumarato, sin embargo, no es capaz de producir p-coumarato a partir de tirosina, por lo que requiere de una enzima heteróloga.

4.2 Espacio metabólico extendido de los organismos de BioCyc

Para *Escherichia coli* que es el organismo modelo más utilizado en biotecnología para la producción de metabolitos heterólogos se encontró que puede producir 1,093 compuestos de forma nativa y 3,268 compuestos más si se le introducen genes externos para formar rutas con 18 reacciones o menos, estos valores son mayores a los calculados por Carbonell y colaboradores en 2011, donde mencionan que *E. coli* produce 966 compuestos y puede producir 2,338 compuestos más, lo

que con el algoritmo desarrollado aquí significa un potencial aumento del 40% en la cantidad de metabolitos que se pueden producir en esta bacteria. El metabolito que requiere más reacciones para producirse (vía más larga) partiendo de los compuestos que produce naturalmente *E. coli* es el 10-(metiltio)-2-oxodecanoata que está a 18 reacciones lineales, la **Figura 16** muestra la estructura química de este compuesto. Estos valores son similares a los encontrados por Chatsurachai y colaboradores en 2012 que son 3,244 potenciales compuestos heterólogos producidos, sin embargo, ellos encuentran que la distancia al compuesto más lejano es de 33 reacciones. Es decir, que nuestro algoritmo pudiera optimizar el encontrar vías metabólicas más cortas, porque permite producir todos los metabolitos utilizando sólo 18 reacciones y aunque se permitan más reacciones no se aumentaría la cantidad de metabolitos producidos porque ya todos fueron conectados.

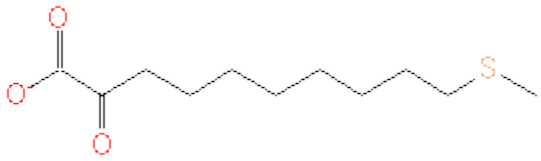
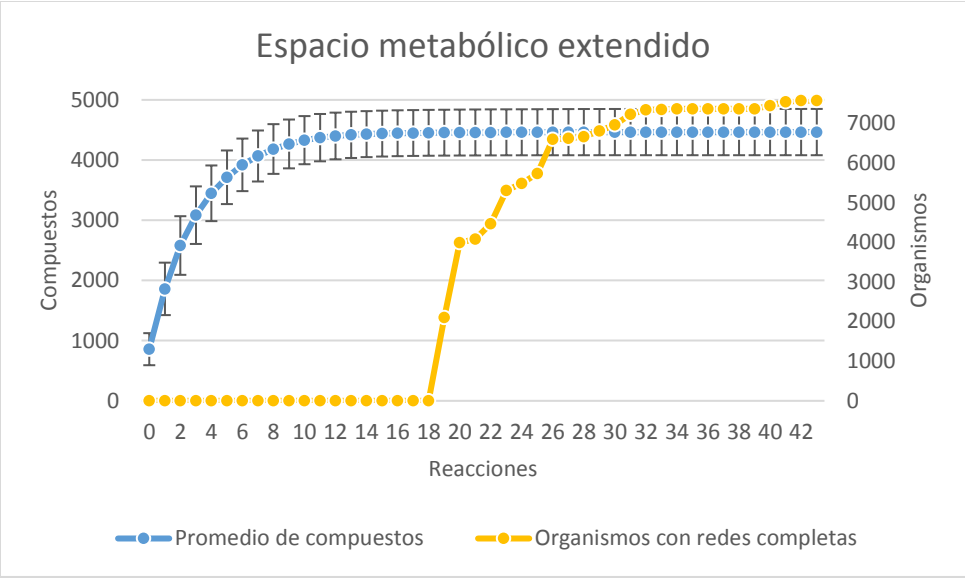


Figura 16. Estructura química del 10-(metiltio)-2-oxodecanoato (CPDQT-41 en MetaCyc).

Este compuesto es el que requiere una ruta más larga de síntesis para *E. coli*, con 18 reacciones lineares, es intermediario en la ruta de síntesis de glucosinolatos en la ruta PWYQT-4450 de MetaCyc, estos compuestos dan el sabor caracteristico a las brasicáceas.

De manera general, considerando para todos los organismos, se encontró que los compuestos más lejanos estaban a 43 reacciones de los metabolitos producidos por el organismo huésped, en la **Gráfica 2** se muestra el promedio de cuantos compuestos pueden producir todos los organismos de la base de datos utilizando rutas con distintos números máximos de reacciones. El organismo que pudiera tener el espacio metabólico más extendido es *Mycobacterium parascrofulaceum* ATCC BAA-614 con 5,618 potenciales compuestos que puede producir de manera heteróloga. Mientras que el organismo que produce una mayor cantidad de metabolitos de manera nativa es *Mycobacterium smegmatis* MC2 155 con 1725 compuestos. Esto es coherente ya que pertenecen a la clase Actinobacteridae,

bacterias reconocidas por su una gran potencial y diversidad metabólica y de las cuales se extraen antibióticos y compuestos para la industria farmacéutica (Lee *et al*., 2014).



Gráfica 2. Promedio del tamaño del espacio metabólico extendido de todos los organismos contenidos en BioCyc añadiendo distintos números de reacciones.

Para cada organismo se expandió su red metabólica hasta 43 reacciones, que es la distancia al compuesto más lejano. Para cada distancia se calculó el promedio de compuestos que produce cada organismo y se graficó junto a su desviación estándar (Barras de error). También se graficó el número de organismos que ya no pueden crecer su red metabólica más allá de determinado número de reacciones.

4.3 Análisis de la red generada por las reacciones de BioCyc

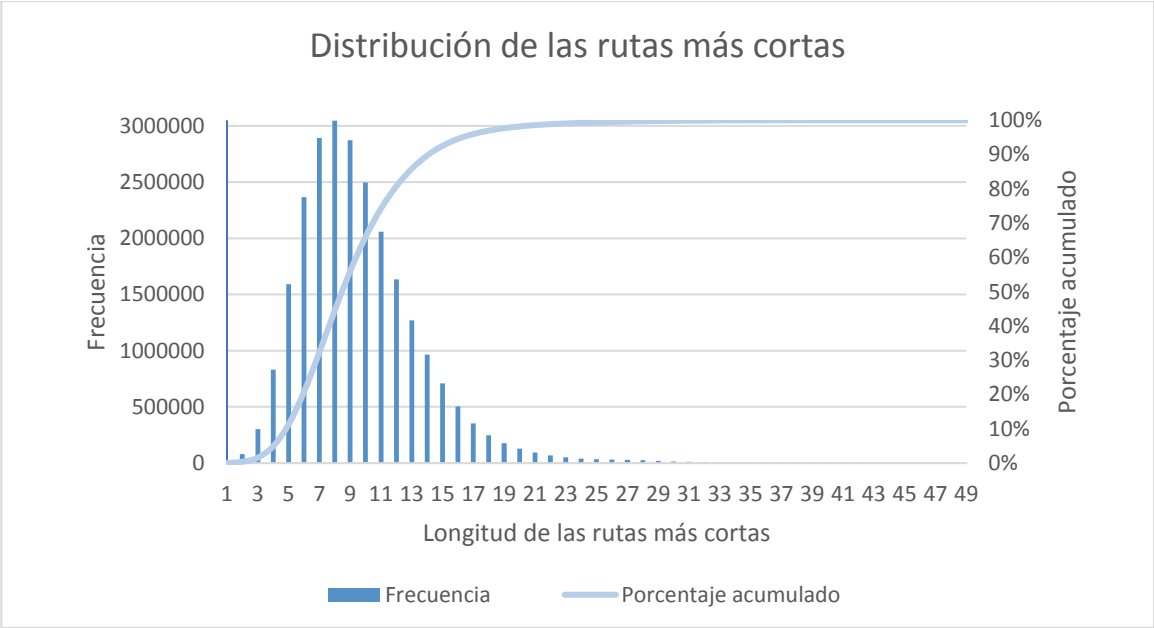
A partir de la red generada con las reacciones de BioCyc (**Figura 17**) y representada como un grafo se utilizó Cytoscape para calcular algunos parámetros de conectividad de la red y se encontró que existen 396 componentes conectados, que representan a grupos de compuestos que forman un grupo de compuestos conectados entre ellos, pero no conectados con otros grupos, con un componente principal que es el que contiene a la mayoría de los compuestos enlazados. Como ejemplos de algunas rutas y reacciones aisladas incluyen la ruta de degradación del 2,4,6-trinitrotolueno o la ruta de biosíntesis de la escopolamina.

La ruta más corta de un compuesto dado a cualquier otro compuesto es el camino que los conecta con el menor número de aristas, para un par de compuestos pueden existir más de una ruta más corta si es que más de un camino llega utilizando el mismo número de aristas. El número de rutas más cortas es calculado con el algoritmo de búsqueda amplia (breadth-first search en inglés) o alguno derivado de su forma estándar, el funcionamiento de este algoritmo es el siguiente (Newman, 2001):

1. Se toma un nodo y se le asigna una distancia d = 0.
2. Para cada nodo con distancia d se buscan los nodos conectados a él.
3. Si un nodo del paso anterior ya tenía asignado una distancia, el nodo es ignorado en el siguiente paso.
4. Se le asigna un valor de d = d + 1 a cada nodo recuperado del paso 2.
5. Se repiten el proceso desde el paso 2 hasta que no quedan nodos sin asignar un valor o ya no es posible conectar más nodos.

Tomando en cuenta lo anterior las rutas más cortas indican cual es la ruta con la menor distancia expresada en número de aristas que conecta dos compuestos dados, ignorando todas las rutas que tengan un mayor número de aristas. Para la red de reacciones de BioCyc las rutas más cortas encontradas es de 24’989,388 que representa el 33.56% de las 74’459,641 combinaciones posibles, la distribución de frecuencia se muestra en la **Gráfica 3**. El algoritmo busca la ruta más corta para conectar cualquier compuesto de la red con cualquier otro compuesto, esto significa que en promedio; a partir de cada compuesto de la red se puede producir un tercio del total de compuestos de la red. Teniendo en promedio que recorrer 9.63 aristas (reacciones químicas) para poder lograrlo. Observamos que el diámetro de la red es de 49 aristas, esto significa que los compuestos más alejados se encuentran a una distancia de 49 aristas o reacciones enzimáticas. Se debe evitar confundir estos valores con las distancias entre los compuestos producidos por un organismo y algún metabolito heterólogo, ya que estas distancias se calculan para todos los nodos de la red de reacciones de BioCyc lo que incluye distancias de metabolitos internos a internos y de metabolitos externos a externos, la cantidad de reacciones

promedio para alcanzar un metabolito heterólogo a partir de un compuesto producido por el organismo varía dependiendo del organismo y será menor a las mostradas para la red general.



Gráfica 3. Frecuencia de las rutas más cortas para conectar dos compuestos cualesquiera en la red de reacciones de BioCyc.

El eje X indica la cantidad mínima de reacciones que se requieren para conectar dos compuestos, mientras que el eje Y indica la frecuencia con que ocurren estas rutas mínimas, en el eje Y secundario se indica la frecuencia en porcentaje acumulado y se observa que el 95% de las rutas se encuentran con un máximo de 17 reacciones.

A pesar de que el diámetro de la red es de 49 aristas, de acuerdo con la distribución de frecuencias, el 95% de las rutas más cortas tienen como máximo una distancia de 17 aristas, como se aprecia en la **Gráfica 3**. Este dato es importante para optimizar el tiempo en la búsqueda de rutas metabólicas, ya que se puede establecer como el máximo número predeterminado de reacciones a 17, y con este valor se estarían considerando a casi todas las rutas para la producción de metabolitos, sin embargo, si así se desea se puede cambiar el máximo número de reacciones a un número mayor en el programa para buscar rutas más con solo cambiar un menú desplegable. Además en los organismos, al poseer un mayor

número de precursores es posible tener varias rutas para producir un metabolito, y de entre estas rutas se puede escoger la de menor distancia.

4.3.1 Cálculo de los índices de centralidad

La centralidad de la carga es similar a la centralidad intermedia, sin embargo, esta última se calcula únicamente para las rutas más cortas en las que interviene un compuesto a evaluar, para calcular la centralidad intermedia se suma para cada par de compuestos la cantidad de rutas más cortas que pasan por el compuesto del cual se quiere calcular la centralidad entre la cantidad de rutas más cortas que existen. Así, un compuesto que únicamente sirve como intermediario entre otros dos compuestos que tienen sólo una ruta más corta tendrá una centralidad intermedia de 1, porque sólo habrá una ruta disponible y ese compuesto está presente en esa ruta, ignorando todas las demás rutas existentes, es por esta razón que esta centralidad no es calculada porque no ofrece información relevante para este trabajo. Con la centralidad de cercanía se tiene un problema similar al de la centralidad intermedia en el cual sólo es calculado para las rutas más cortas conectadas, sin embargo, esta centralidad indica que tan cerca está un compuesto de todos los compuestos a los que se conecta y podemos obviar el problema de la conexión entre los compuestos si la calculamos para los compuestos con mayor centralidad de la carga.

Para determinar cuáles fueron los compuestos que eran intermediarios en un mayor número de las rutas más cortas, se calculó la centralidad de la carga, este parámetro puede ayudar a encontrar vértices (metabolitos) importantes en la red, ya que entre mayor sea su valor (número de rutas en la que participa) es más probable que participe como intermediario en cualquier ruta elegida. Puede darse el caso en que a pesar de estar presentes en rutas más cortas, estas rutas no necesariamente estén presentes en la naturaleza o sean usadas de manera nativa por un organismo.

Para calcular los índices de centralidad se utilizó NetworkX (Hagberg, Schult, & Swart, 2008) esto debido a que Cytoscape genera errores en el cálculo de algunos tipos de centralidad para multígrafos porque al llegar a las aristas múltiples genera

rutas redundantes y ciclos (Anónimo, NetworkAnalyzer Settings, 2013). En contraste, NetworkX tiene una implementación para multígrafos dirigidos llamado MultiDiGraph que puede manejar sin problema este tipo de grafos y que utiliza propiedades internas de la red para calcular la centralidad de la carga en lugar de contar cada ruta (Newman, 2001).

De la red de reacciones de BioCyc se recuperaron los 10 compuestos con mayor carga y, se calculó la centralidad de cercanía y compiladas en la **Tabla 8**. Estos compuestos se consideran como los posibles precursores para una gran diversidad de metabolitos y pueden considerarse como una meta de optimización para aumentar su producción mediante ingeniería metabólica, esto para incidir positivamente en aumentar la producción de varios compuestos que se deseen sintetizar a partir de estos. Por ejemplo, el compuesto (2*E*,6*E*)-farnesil difosfato, tuvo una de las centralidades de carga grande, debido a que es un intermediario en la ruta del esterol y funciona como precursor de los sesquiterpenos que son la clase más diversa de isoprenoides en la cual se han caracterizado más de 7,000 compuestos (Asadollahi, Maury, Schalk, & Clark, 2009). Utilizando ingeniería metabólica es posible aumentar la producción de este compuesto en distintos organismos y podrían ser de utilidad en la producción de un gran número de compuestos que se derivan de él, esto permite pensar en modularizar las modificaciones genéticas más comunes para mejorar la producción de metabolitos.

Tabla 8. Valores de centralidad para los principales componentes de la red metabólica de BioCyc.

ID de BioCyc	Nombre	Centralidad	
		Carga	Cercanía
ACETYL-COA	Acetil-Coenzima A	0.11099	0.11894
PYRUVATE	Piruvato	0.10791	0.11959
MALONYL-COA	Malonil-Coenzima A	0.04187	0.11477
TYR	L-Tirosina	0.04159	0.11379
ACET	Acetato	0.03924	0.11145
CPD-4211	Pirofosfato de dimetilalilo	0.03781	0.10059
DELTA3-ISOPENTENYL-PP	Isopentenil difosfato	0.03292	0.09703
CPD-16653	4-*O*-dimetilalil-L-tirosina	0.02532	0.10615
FARNESYL-PP	(2*E*,6*E*)-farnesil difosfato	0.02411	0.09869
FORMATE	Formato	0.02404	0.10637

Figura 17. Grafo que representa las reacciones presentes en BioCyc.

A partir de las reacciones de BioCyc se generaron pares reactantes ignorando los compuestos altamente conectados, estos pares reactantes se utilizaron para crear un multígrafo dirigido. Los nodos representan compuestos y las aristas representan pares reactantes. La imagen fue creada utilizando Cytoscape. Cada grupo de compuestos conectados entre sí, pero aislados del resto es llamado un componente de la red, siendo el componente principal el que tiene el mayor número de nodos conectados.

5 Discusión

La identificación y enumeración de rutas metabólicas permitió reconstruir el 77% de las rutas presentes en la clase de Metabolitos Secundarios de BioCyc. Las rutas no encontradas se debieron a que BioCyc utiliza compuestos genéricos para describir reacciones, por ejemplo, los aminoácidos son un compuesto genérico que incluye todos los compuestos químicos con un grupo amino y un grupo carboxílico, esta clasificación incluye incluso a aminoácidos no proteicos. El problema con esta clasificación es cuando en las reacciones se genera un compuesto genérico a partir de otro compuesto genérico, por ejemplo, en la reacción ACYL-COA-HYDROLASE-RXN de MetaCyc (**Figura 18**) los compuestos que puede contener Acil-CoA son 162, mientras que los Carboxilatos incluyen 1,157, entonces la cantidad de pares reactantes que se formarían de estos dos compuestos en esta reacción son 187,434, de los cuales escasamente uno sería correcto para una reacción específica. BioCyc no brinda información ni herramientas para poder generar o corregir estas anotaciones. Se podrían rescatar algunos pares reactantes útiles a partir de estas reacciones, por ejemplo, cuando existan dos pares reactantes y uno de ellos sea el producto de la conversión de Acil-CoA a Carboxilatos, y si el otro par reactante no contiene compuestos genéricos, esta reacción puede ser rescatada si se hace una verificación manual de la reacción involucrada. Este caso ocurre en la reacción RXN-15375 de MetaCyc donde el 3-sulfinopropionato se convierte en 3-sulfinopropanoil-CoA, sin embargo, no es posible saber si esta reacción puede llevarse a cabo en todos los organismos debido a que la reacción podría no llevarse a cabo con todas las instancias de ese compuesto genérico.

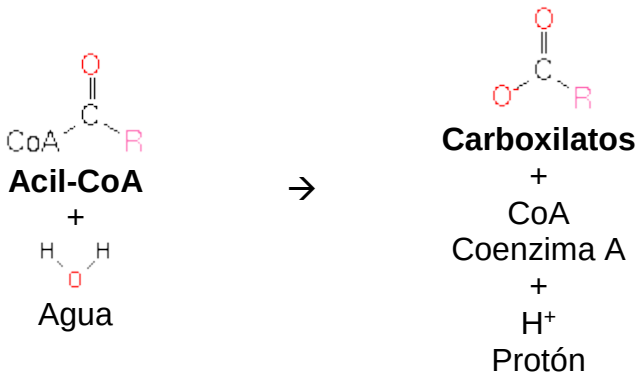


Figura 18. Reacción para la producción de carboxilatos.

Se encontró una gran redundancia en las rutas metabólicas debido a la presencia de reacciones redundantes, en las cuales los metabolitos principales se conservan, sin embargo, los metabolitos transportadores de grupos funcionales varían, por lo que las clasificamos como reacciones diferentes. Eliminar estas reacciones redundantes no es una tarea trivial, se tendría que estudiar detalladamente cada grupo de casos. Por ejemplo, en el caso de un compuesto genérico donador de electrones en ocasiones parece ser substituido por el oxígeno, ver **Figura 14**. Se podría pensar que dicha reacción es la misma ya que el oxígeno podría funcionar como donador de electrones, pero también nos podemos dar cuenta de que las energías de reacción son distintas, lo que podría indicar un mecanismo de reacción distinto, o la no equivalencia entre los dos compuestos, o bien una desviación en el cálculo de las energías de reacción. Por estas razones, aún no es claro si estas reacciones deben considerarse como la misma o no. La presencia de rutas redundantes no es un problema cuando el número de rutas encontradas son pocas ya que es posible reconocerlas revisando cuales son los compuestos principales en cada reacción, sin embargo, de forma general el número de rutas crece exponencialmente conforme crece el número de reacciones máximas permitidas y aunado a esto el tener reacciones redundantes también hace crecer exponencialmente la cantidad de rutas predichas.

En BioCyc existen reacciones compuestas definidas como reacciones que se pueden descomponer en otras reacciones más simples. Por ejemplo, la reacción RXN-14306 está formada por las reacciones 4.2.1.111-RXN y RXN-8986, esta reacción es descrita en la **Figura 3**. Las reacciones compuestas no siempre están especificadas como compuestas (SRI International, 2016), por lo que se pueden producir errores al momento de medir el costo de las transformaciones químicas, además de generar rutas redundantes. Para evitar estas redundancias es necesario verificar que todas las sub-reacciones de una reacción compuesta esté presente en los registros, ya que nuestro algoritmo es capaz de procesarlas, sin embargo, no puede hacerlo si no se le indica que son compuestas.

Se intentó basar el ordenamiento de las rutas de acuerdo con parámetros cinéticos de las enzimas utilizadas en ellas, sin embargo, solamente existe información para el 0.042% de las enzimas recuperadas en la base de datos local, por lo que no contribuían realmente a la evaluación de las rutas. Es por lo que se optó por hacer una evaluación termodinámica, la cual se complementa con la evaluación del costo de las transformaciones químicas de los productos con respecto a los precursores.

Por otro lado, los valores cinéticos que se recuperaron de Uniprot no fueron todos los disponibles, debido a que los compuestos de Uniprot no poseen el mismo nombre que en BioCyc o algún enlace entre ambas bases de datos, así que se intentaron recuperar todos los posibles con sus sinónimos. Se debe considerar crear una unificación en el nombre de los compuestos de las dos bases de datos de forma manual, aunque aún con esto no se lograría recuperar una cantidad significativa de datos de las enzimas para utilizarlos como parámetro de evaluación, ya que los valores que no se agregaron son 756. Algunos de los registros no agregados hacen referencia al valor de Km para cofactores o transportadores de grupos funcionales o energía como NAD. Este problema para encontrar el mismo compuesto en dos bases de datos es discutido por Almant, *et al*. en 2013 sobre los esfuerzos por unificar KEGG y BioCyc, sin que se haya encontrado una solución definitiva.

A pesar de que la determinación de la factibilidad termodinámica de una ruta es de gran utilidad, se debe de tener en cuenta que la mayoría de los valores termodinámicos usados en la base de datos fueron generados computacionalmente y no son completamente precisos (Noor, Haraldsdíttir, Milo, & Fleming, 2013) (Jankowski, Henry, Broadbelt, & Hatzimanikatis, 2008). Por lo que en ocasiones las rutas encontradas pueden ser clasificadas como no posibles termodinámicamente a pesar de existir en las rutas de referencia. También se debe tomar en cuenta que las concentraciones que los metabolitos pueden alcanzar se establecieron de forma genérica para la mayoría de ellos, excepto para algunos cuantos que pudieron medirse con mayor exactitud como la concentración de H, Agua, Co-A y otros, por esta razón, podría ser que las concentraciones reales de los compuestos pudieran diferir de las establecidas y no reflejar el valor real.

La información introducida a MetaCyc o actualizada puede ser propagada a los demás organismos. Durante la generación de la base de datos local se encontró que algunas reacciones, enzimas y genes involucrados en la síntesis de carotenoides no estaban presentes en *Pantoea ananatis*, a pesar de que en MetaCyc se indicaba que estas reacciones si estaban presentes en este organismo. Parece ser que estas reacciones fueron introducidas a MetaCyc manualmente y sólo se hizo su asignación en estos organismos utilizando su identificador taxonómico. Tratamos de propagar estas anotaciones a los organismos a los que pertenecen, pero es probable que no se hayan podido expandir todas las existentes debido a la falta de identificadores taxonómicos.

Se debe de tomar en cuenta que las anotaciones de los genomas se realizaron con distintos programas y por lo tanto existen variaciones y errores en los genes anotados y la asignación de funciones, por ejemplo, como se menciona anteriormente en la base de datos existen genes de hasta 2.5 millones de pares de bases.

Es necesario crear reglas bien definidas para determinar si un organismo puede producir un compuesto, también se debe definir bajo que condiciones de cultivo el organismo puede producir estos compuestos, debido a que la forma actual en la cual las bases de datos agregan las anotaciones de compuestos a los organismos genera problemas en los algoritmos de reconstrucción de rutas metabólicas. Una forma de estandarizar las anotaciones es verificar si los compuestos predichos pueden ser producidos a partir del metabolismo central, sin embargo, esto requiere conocer todas las reacciones presentes en un organismo. Otra opción es definir a partir de que precursor se quiere producir un compuesto deseado.

En MetaCyc existen reacciones que fueron introducidas manualmente y que no pertenecen a organismos presentes en BioCyc. Estas reacciones fueron tratadas de unificarse con otras bases de datos. La mayoría de las enzimas para estas reacciones tienen una liga para consultarlas en Uniprot, en donde se pueden recuperar sus secuencias de aminoácidos y otros valores, pero no las secuencias

de ADN de los genes que las codifican, por esta razón se recurrió a KEGG, que si tiene secuencias para los genes de las enzimas.

Para crear nuestra base de datos local escribimos un programa capaz de procesar la información de BioCyc para todos los organismos, durante el desarrollo de este programa hicimos énfasis en que fuera lo más autónomo posible, esto porque BioCyc es una colección que está siendo constantemente actualizada y cada cierto tiempo libera nuevas versiones, así nosotros también podemos actualizar nuestra base de datos local sin mucho esfuerzo.

Se ha considerado añadir información de otras bases de datos a nuestra base de datos local, sin embargo, muy probablemente esto generaría redundancia en los registros agregados, debido a que no es sencillo encontrar equivalencias entre compuestos y reacciones presentes en dos bases de datos generadas independientemente. También se debe considerar que las filosofías o usos de las distintas bases de datos difiere por lo que pudieran no ser compatibles.

Si bien nuestro algoritmo de reconstrucción de rutas metabólicas y el ordenamiento de estas es completamente funcional, nos vemos limitados por la exactitud y disponibilidad de la información presente en BioCyc y otras bases de datos. Sin embargo, durante las pruebas realizadas sobre nuestro programa pudimos darnos cuenta de errores presentes en BioCyc que pudieran pasar desapercibidos para la mayoría de los usuarios, pero que aquí fueron fácilmente reconocidos ya que se le asigna un significado mayor a cada registro, más que ser estadísticas para un organismo. Este tipo de observaciones podría ayudar a BioCyc a recibir más observaciones por parte de los usuarios.

Consideramos que como todo programa el nuestro puede ser mejorado y seguramente así será, sin embargo, es más urgente disponer de bases de datos especializadas en reconstrucción de rutas metabólicas validadas experimentalmente y estén completamente enfocadas a ello, porque existen cuestiones técnicas que no han sido abordadas por las bases de datos multipropósito como BioCyc o KEGG, una de estas cuestiones ha sido discutida anteriormente y radica en crear la definición exacta de que compuestos considerar

como propios de cada organismo. Esperamos poder contribuir a esto al liberar nuestro programa ya que su uso nos permitiría saber cuáles son las rutas de mayor interés para los usuarios y enfocar nuestros esfuerzos a curar estas rutas.

A pesar de todas las limitaciones encontradas durante este proyecto ponemos a disposición un algoritmo y una base de datos que permite reconstruir hiperrutas metabólicas para producir compuestos partiendo de los compuestos producidos por una gran cantidad de organismos no modelo, además de ordenarlas de acuerdo con sus propiedades cinéticas y costo de las transformaciones químicas.

6 Conclusiones

Utilizando la información de BioCyc creamos una base de datos local con información para la reconstrucción de rutas metabólicas, para esto desarrollamos nuestro propio programa en el cual se crean rutas metabólicas a partir de hipergrafos, validando que todos los precursores de cada reacción puedan ser sintetizados por el organismo o se sinteticen en la propia ruta.

Utilizando parámetros termodinámicos para cada reacción determinamos la factibilidad y eficiencia de cada ruta, además del costo de la transformación química, lo que nos permite ordenarlas de acuerdo con cuáles son las que tienen un mayor potencial para la producción de metabolitos. Al final de este proceso obtenemos los genes involucrados de cada ruta y permitimos a los usuarios editarlos. Si bien el algoritmo para la reconstrucción de rutas funciona de acuerdo con lo esperado, aún existen errores y falta información en las bases de datos que requieren ser abordados desde otros puntos de vista para solucionarlos.

Con el uso de nuestro programa esperamos crear una lista de las rutas más buscadas por los usuarios, para curarlas manualmente y brindar una mayor confiabilidad de los resultados. Además, se puede crear una base de datos complementaria con módulos de optimización para la producción de metabolitos nativos de organismos de acuerdo a la lista de compuestos con mayor centralidad de la carga (ver **Tabla 8**), esto permitiría aumentar la producción de una gran cantidad de compuestos debido a la gran cantidad de rutas en las que participan los compuestos con alta carga.

Es importante generar normas más rigurosas para definir que compuestos es capaz de producir un organismo y cuáles no, además de las condiciones en que lo hace, esto para crear una mayor fiabilidad en los precursores seleccionados para la producción de compuestos de interés.

7 Perspectivas

Es necesario curar manualmente las reacciones encontradas en BioCyc, para eliminar las reacciones redundantes y homogenizar la anotación llevada a cabo en la base de datos. Por ejemplo, para el caso de las reacciones químicas, el estándar es indicar el coeficiente estequiométrico de los metabolitos, sin embargo, existen reacciones donde tienen un metabolito repetido en lugar de indicarlo con un número en el coeficiente. Otro ejemplo es cuando las reacciones tienen protones y hay reacciones idénticas sin indicar este protón indicado y, por último, cuando las reacciones tienen componentes genéricos y una secuencia de reacciones están repetidas con instancias de esos componentes genéricos.

Con los pares reactantes de BioCyc verificados manualmente sería posible simular las redes metabólicas dentro de los organismos, y con esto poder indicar cuál es la fuente de carbono añadida al medio de cultivo y de la cual se debe producir el metabolito, permitiendo simular rutas globales más cortas.

Como BioCyc basa sus anotaciones en la presencia de secuencias similares a otras secuencias de enzimas es posible que en ocasiones se anoté algún gen cuyo promotor este inactivo y por lo tanto a pesar de estar presente no sea funcional, desencadenando en que se predigan compuestos que no existen. Para mejorar el uso de los distintos organismos sería importante agregar una interfaz para que los usuarios creen organismos locales personalizados, con esto se podría tomar un organismo de la base de datos y crear una copia de él en la cual se puedan eliminar manualmente compuestos que el usuario considere que no son de importancia o que él haya verificado que no existen en este organismo, también se podrían agregar compuestos que el organismo si produzca y no estén anotados por que aún no se conocen las enzimas que lo producen. Con estas bases de datos locales se puede también generar nuevos organismos desde cero o crear información para cepas similares a algún organismo presente en la base de datos. Esto podría ser de interés para laboratorios que tengan su cepa modelo y en la cual realicen cotidianamente modificaciones genéticas. Además de los organismos, se podría implementar herramientas para definir la constitución de los medios de cultivo para,

de esta manera, poder añadir precursores como suplemento en el medio de cultivo y no sería necesario partir de los compuestos producidos por el organismo.

A partir de los metabolitos que se encontraron con mayor centralidad de la carga se debería de generar una base de datos de modificaciones genéticas para distintos organismos que permitan aumentar su producción de forma más sencilla. Incluso podría crearse una colección de organismos con estas modificaciones que permita modularizar la producción de metabolitos.

Para este proyecto se utilizó una base de datos basada en SQL, sin embargo, la cantidad de registros en cada tabla es tan grande que retarda el cálculo de rutas, por esta razón es importante buscar una alternativa para optimizar el tiempo de respuesta de estas bases de datos tan grandes. Una alternativa podría ser el uso de mongoDB que es una base de datos noSQL con un más rápido desempeño (MongoDB, Inc., 2017).

La cantidad de información almacenada en la colección BioCyc es tan grande que sería impensable curarla manualmente, además de la carencia de información experimental para hacerlo. Por esta razón, la curación de la base de datos basada en los usuarios sería una alternativa que beneficiaría a una mayor cantidad de usuarios repartiendo el esfuerzo para su curación. Por ejemplo, para determinados organismos se podrían ir curando manualmente las rutas más usadas, labor a la que incluso los usuarios podrían contribuir mediante la aprobación o desaprobación de cada ruta o añadiendo comentarios, creando así una red de colaboración que beneficie a todos.

Los genomas incorporados a BioCyc utilizan información de otras bases de datos para ser anotados, pero estas anotaciones no son actualizadas a pesar de que las bases de datos utilizadas para anotar se hayan actualizado para corregir errores. Por esta razón es importante que los usuarios comprueben la anotación de los genes antes de utilizarlos, esto se puede lograr buscándolos en bases de datos como las de Pfam (Finn *et al*., 2015) para encontrar si el gen anotado tiene similitud a otros genes que realizan la reacción que se les atribuye. Esto puede ser realizado dentro del programa utilizando la API de Pfam.

8 Glosario

A

Aristas
En teoría de grafos, representan la interacción entre dos nodos., 17, 18, 20, 52, 66, 67, 68, 69, 71

ATP
Abreviación para Adenosín Trifosfato, es una molécula utilizada para almacenar y transportar energía para reacciones químicas dentro de la célula., 17, 18, 49, 50, 52, 88

B

BioCyc
Colección de bases de datos que almacena información sobre organismos y rutas metabólicas., 5, 7, 10, 11, 12, 13, 14, 17, 30, 31, 32, 33, 34, 36, 37, 38, 47, 48, 49, 53, 54, 55, 56, 58, 62, 64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 76, 78, 79, 80, 90

Biología Sintética
La Biología Sintética es el nombre que recibe una disciplina científica emergente que integra conocimientos de biología, genética, química, ciencia computacional e ingeniería y que tiene como objetivo principal la producción de formas de vida nuevas o mejoradas., 5

C

Catalizadores
Sustancia que acelera una reacción sin participar en ella., 9

Centralidad de cercanía
Índica que tan cercano esta un nodo a todos los nodos para los cuales existe una ruta, entre mayor sea este índice más cercanía hay con los nodos., 19

Centralidad de la carga
Índica cuantas de todas las rutas más cortas pasan por un nodo., 20

Centralidad intermedia
Es la suma de la fracción de todas las rutas más cortas que conectan dos nodos y que pasan por el nodo a evaluar., 19

Compuestos genéricos
Compuestos que representan a un grupo de compuestos., 32, 35, 72, 89

Constante de Michaelis-Menten. *Véase* Km

Costo de las transformaciones químicas
Indica cuantos átomos y enlaces fueron cambiados durante las reacciones presentes en una ruta., iii, 45, 47, 73, 74, 77

Cytoscape
Programa para análisis y visualización de redes., 30, 53, 66, 69, 71

D

Distancia Manhattan
Indica la distancia entre dos puntos medida por las diferencias absolutas de sus coordenadas., 45, 46, 47

E

EcoCyc
Base de datos con información referente a *E. coli* MG1655., 11, 21, 36

Enzimas
Una enzima es una proteína que cataliza las reacciones bioquímicas del metabolismo., 8, 9, 11, 14, 15, 30, 33, 36, 37, 38, 44, 47, 48, 54, 55, 56, 58, 59, 61, 63, 74, 75, 76, 79

Espacio metabólico
Conjunto de reacciones y compuestos presentes en un organismo., 5, 14, 21, 39, 48, 65

Espacio metabólico extendido
Conjunto de reacciones y compuestos que se añaden a un organismo., 14, 21, 39, 48, 65

Estado de equilibrio de la reacción
Estado en el que para una reacción reversible, las dos direcciones de la reacción ocurren con la misma velocidad., 10

G

Grafos
Representación simbólica de los elementos constituidos de un sistema o conjunto, mediante esquemas gráficos., 16, 17, 18, 19, 20, 21, 69

H

Hipergrafos
Es un tipo de grafo en el cual una arista puede unir a más de dos nodo., 5, 20, 21, 78

I

Índices de centralidad

S

V

X

9 Bibliografía

Chae, T. U., Kim, W. J., Choi, S., Park, S. J., & Lee, S. Y. (2015). Metabolic engineering of Escherichia coli for the production of 1,3-diaminopropane, a three carbon diamine. *Scientific Reports, 5*, 13040.

Carbonell, P., Fichera, D., Pandit, S., & Faulon, J.-L. (2012). Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Sysmtems Biology, 6*(10), 1-18.

Carbonell, P., Planson, A.-G., Fichera, D., & Faulon, J.-L. (2011). A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Systems Biology, 5*(122), 1-18.

Carbonell, P., Parutto, P., Herisson, J., Pandit, S. B., & Faulon, J.-L. (2014). XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Research, 42*, W389-W394.

Larson, R., & Falvo, D. C. (2017). *Elementary Linear Algebra.* Boston: Houghton Mifflin Harcourt Publishing Company.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C., Keseler, I., . . . Karp, P. (2016). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research, 44*(1), D471-80.

Chatsurachai, S., Furusawa, C., & Shimizu, H. (2012). An in silico platform for the design of heterologous pathways in nonnative metabolite production. *BMC Bioinformatics, 13*, 93-104.

Lee, L.-H., Zainal, N., Azman, A.-S., Eng, S.-K., Goh, B.-H., Yin, W.-F., . . . Chan, K.-G. (2014). Diversity and Antimicrobial Activities of Actinobacteria Isolated from Tropical Mangrove Sediments in Malaysia. *The Scientific World Journal*, 1-14. doi:http://dx.doi.org/10.1155/2014/698178

Li, C., Henry, C. S., Jankowski, M. D., Ionita, J. A., Hatzimanikatis, V., & Broadbelt, L. J. (2004). Computational discovery of biochemical routes to specialty chemicals. *Chemical Engineering Sciense, 59*, 5051-5060.

Loescheke, A., & Thies, S. (2015). Pseudomonas putida-a versatile host for the production of natural products. *Appl Microbiol Biotechnol, 99*, 6197-6214.

Altman, T., Travers, M., Kothari, A., Caspi, R., & Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics, 14*, 112-127.

Anónimo. (20 de Junio de 2013). *NetworkAnalyzer Settings*. Recuperado el 3 de Junio de 2017, de NetworkAnalyzer: http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/

Anónimo. (13 de Febrero de 2017). *Pathway Tools Question & Answer forum for SRI's Pathway Tools*. Recuperado el 20 de Mayo de 2017, de SRI's Pathway Tools: https://ask.pathwaytools.com/question/17956/km-from-flat-files/

Apache Commons. (28 de Agosto de 2016). *Commons Math: The Apache Commons Mathematics Library*. Recuperado el 2 de 04 de 2017, de Apache Commons: http://commons.apache.org/proper/commons-math/

Asadollahi, M. A., Maury, J., Schalk, M., & Clark, A. (4 de Enero de 2009). Enhancement of Farnesyl Diphosphate Pool as Direct Precursor of Sesquiterpenes Through Metabolic Engineering of the Mavalonate Pathway in Saccharomyces cerevisiae. *Biotechnology and Bioengineering, 106*(1), 86-96.

Berg, J., Tymoczko, J., & Stryer, L. (2002). *Biochemistry* (Quinta ed.). New York: W H Freeman.

BioCyc. (2016). *INTRODUCTION TO BIOCYC*. Recuperado el 14 de 06 de 2017, de BioCyc Database Collection: https://biocyc.org/intro.shtml

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., . . . Xenarios, I. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. En D. Edwards, *Plant Bioinformatics Methods and Protocols* (págs. 23-54). Perth , Australia: Springer.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology, 25*(2), 163-177.

Draths, K. M., & Frost, J. W. (1995). Environmentally Compatible Synthesis of Catechol from D-Glucose. *J. Am. Chem. Soc., 117*, 2395-2400.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchel, A. L., . . . Bateman, A. (15 de Diciembre de 2015). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research, 44*(D1), D279-D285.

Gandía-Herrero, F., & García-Carmona, F. (2014). Escherichia coli protein YgiD produces the structural unit of plant pigments betalains: characterization of a prokaryotic enzyme with DOPA-extradiol-dioxygenase activity. *Appliedl Microbiology and Biotechnology, 3*, 1165-1174.

Goh, K.-I., Kahng, B., & Kim, D. (31 de Diciembre de 2001). Universal Behavior of Load Distribution in Scale-Free Networks. *Physical Review Letters, 87*(27).

GraphML team. (11 de Julio de 2016). *The GraphML File Format*. Recuperado el 20 de Mayo de 2017, de Graph Drawing: http://graphml.graphdrawing.org/

Guo, A. C., Jewison, T., Wilson, M., Liu, Y., Knox, C., Djoumou, Y., . . . Wishart, D. S. (2013). ECMDB: The E. coli Metabolome Database. *Nucleic Acids Research, 41*, D625-D630.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. En G. Varoquaux, T. Vaught, & J. Millman (Ed.), *Proceedings of the 7th Python in Science Conference (SciPy2008)*, (págs. 11-15). Pasadena.

Handorf, T., Ebenhöh, O., & Heinrich, R. (2005). Expanding Metabolic Network: Scopes of Compounds, Robustness, and Evolution. *J Mol Evol, 61*, 498-512. doi:DOI: 10.1007/s00239-005-0027-1

Hu, Y., Rolfs, A., Bhullar, B., Murthy, T. V., Cong Zhu, Berger, M. F., . . . LaBaer, J. (2007). Approaching a complete repository of sequence-verified protein-encoding clones for Saccharomyces cerevisiae. *Genome Research, 4*, 536-543.

Jang, H.-J., Yoon, S.-H., Ryu, H.-K., Kim, J.-H., Wang, C.-L., & Kim, J.-Y. (2011). Retinoid production using metabolically engineered Escherichia coli with a two-phase culture system. *Microbial Cell Factories, 10*, 59-61.

Jankowski, M., Henry, C., Broadbelt, L., & Hatzimanikatis, V. (2008). Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophysical Journal, 95*(3), 1487-1499.

Junker, B., & Schreiber, F. (2008). *Analysis of Biological Networks.* New Jersey: Wiley-Interscience.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res, 28*(1), 27-30.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res, 28*, 27-30.

Karp, P. D., Latendresse, M., & Caspi, R. (2011). The Pathway Tools Pathway Prediction Algorithm. *Standards in Genomic Sciences, 5*, 424-429. doi:10.4056/sigs.1794338

Kerp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., . . . López-Bigas, N. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res., 33*(19), 6083-6089.

McShan, D., Rao, S., & Shah, I. (2003). PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics, 19*(13), 1692-1698. doi:doi:10.1093/bioinformatics/btg217

Ma, H., & Zeng, A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics, 19*(2), 270-277.

MongoDB, Inc. (2017). *Reinventando la gestión de datos*. Obtenido de MongoDB: https://www.mongodb.com/es

Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., & Kanehisa, M. (2010). PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Research, 38*, W138-W143.

NetworkX Developers. (2015). *closeness_centrality*. Recuperado el 6 de julio de 2017, de NetworkX: https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.centrality.closeness_centrality.html#networkx.algorithms.centrality.closeness_centrality

NetworkX Developers. (2015). *load_centrality*. Recuperado el 6 de Junio de 2017, de NetworkX: https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.centrality.load_centrality.html

NetworkX Developers. (2015). *betweenness_centrality*. Recuperado el 6 de junio de 2017, de NetworkX: https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.centrality.betweenness_centrality.html#networkx.algorithms.centrality.betweenness_centrality

Newman, M. (2001). Sientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *PHYSICAL REVIEW E, 64*, 1-7.

Noor, E., Bar-Even, A., Flamholz, A., Reznik, E., Liebermeister, W., & Milo, R. (2014). Pathway Thermodinamics Highlghts Kinetic Obstacles in Central Metabolism. *PLOS Computational Biology, 10*(2), e10003483.

Noor, E., Haraldsdíttir, H., Milo, R., & Fleming, R. (2013). Constent Estimation of Gibbs Energy Using Component Contributions. *PLOS Computational Biology, 9*(7), e1003098.

R Core Team. (2013). *A language and environment for statistical.* Obtenido de R project: http://www.R-project.org

SRI International. (2016). *MetaCyc Reactions Class: Composite Reactions*. Recuperado el 3 de Junio de 2017, de BioCyc Database Collection: https://biocyc.org/META/new-image?object=Composite-Reactions

SRI International. (19 de 06 de 2017). *Metabolic Route Search of Escherichia coli K-12 substr. MG1655*. Obtenido de BioCyc: https://biocyc.org/meteng?organism=ECOLI

UniProt Consortium. (2017). *UniProt*. Recuperado el 10 de Julio de 2017, de UniProt: http://www.uniprot.org/

Verhoef, S., Ruijssenaars, H. J., de Bont, J. A., & Wery, J. (2007). Bioproduction of p-hydroxybenzoate from renewable feedstook by solvent-tolerant Pseudomonas putida S12. *Journa of Biotechnology*, 49-56.

8. Apéndices

Apéndice 1. Compuestos descartados de la red metabólica

Tabla 9. Tabla con compuestos altamente conectados y pequeñas moléculas ignorados en la red metabólica.

PROTON	WATER	OXYGEN-MOLECULE	CL-
AMMONIUM	CPD-12575	NITROGEN-MOLECULE	HCL
Oxidized-ferredoxins	NITRATE	Elemental-Sulfur	K+
CO-A	Acceptor	HYDROGEN-MOLECULE	Pi
FAD	FADH2	CARBON-DIOXIDE	PPI
NADP	GTP	NAD	P3I
GMP	AMP	NAD-P-OR-NOP	CA+2
ATP	ADP	NADH-P-OR-NOP	NI+2
HYDROGEN-PEROXIDE	PAPS	NITROUS-OXIDE	CR+3
FMN	CoM	Donor-H2	CR+6
CPD-846	CTP	METHYLENE-THF	ZN+2
3-5-ADP	ARSENATE	TUNGSTATE	SE-2
CPD-678	GDP	SELENITE	F-
SELENATE	THF	CPD-7046	BR-
CPD-387	CPD-12755	UV-Light	FE+3
Reduced-ferredoxins	AMMONIA	CPD-763	FE+2
S-ADENOSYLMETHIONINE	NADPH	CPD-7425	HCO3
CPD-7592	NITRITE	H2CO3	CD+2
CPD-4544	NADH	CPD-12799	HS
NH4OH	NMNH	PHOSPHONATE	SO3
S-ADENOSYLMETHIONINAMINE	SEPO3	CPD-14557	S2O3
ADENOSYL-HOMO-CYS	UDP	GLUTATHIONE	HSCN
GMP	TDP	SULFATE	MG+2
Me-CoM	HSO3	OXIDIZED-GLUTATHIONE	AG+
CMP	UMP	LEU-tRNAs	OH
CPD-12377	HCN	10-FORMYL-THF	CU+2
ITP	TMP	5-METHYLTHIOADENOSINE	CU+
COA-GROUP	CDP	DIHYDROFOLATE	CO+2
PHOSPHATE-GROUP	IDP	IRON-CHELATE	
GLUTATHIONE-SULFITE	TTP	5-METHYL-THF	
GLUTATHIONE-SULFIDE	UTP	FMNH2	

Apéndice 2. Compuestos genéricos agregados a la base de datos.

Tabla 10. Tabla con lista de compuestos genéricos agregados a la base de datos.

Oxidized-ferredoxins	Reduced-ferredoxins	Acceptor	Donor-H2
Deaminated-Amine-Donors	Aminated-Amine-Donors	2-Oxo-Acids	Amino-Acids
Charged-LEU-tRNAs	LEU-tRNAs	ACYL-COA	Carboxylates
NADH-P-OR-NOP	NAD-P-OR-NOP	Charged-TYR-tRNAs	TYR-tRNAs

Apéndice 3. Compuestos en la clase de metabolitos secundarios de BioCyc

Tabla 11. Compuestos en la clase de metabolitos secundarios, en verde los compuestos para los cuales se pudo reconstruir al menos una ruta metabólica y en naranja los que no.

CPD-15211	CPD-17046	CPD-14450	CPD-17454
CPD-15212	CPD-17045	CPD-13661	CPD-10198
CPD-15210	CPD-17043	CPD-13655	CPD-10213
CPD-10214	CPD-16887	CPD-729	3-hydroxy-L-kynurenine
CPD-18443	CPD-16888	CPD-9445	CPD-10211
CPD-10210	CPD-16886	CPD0-2108	N-formylkynurenine
CPD-18438	CPD-16885	CPD0-2106	CPD-10207
CPD-403	CPD-16744	CPD-196	CPD-10201
CPD-18328	CPD-16745	CPD-10206	P-COUMAROYL-COA
CPD-18332	CPD-16742	CPD-13647	CPD-10205
CPD-18331	CPD-16741	CPD-195	CPD-10209
CPD-18333	CPD-16740	COUMARATE	CPD-10208
CPD-18327	CPD-16738	CPD-551	CPD-10204
CPD-45	CPD-16737	CPD-1106	CPD-4588
STIPIT-CPD	CPD-16736	CPD-235	CPD-10179
CPD-18153	CPD-10177	CPD-235	GERANYLGERANYL-PP
CPD-18151	CPD-16727	CPD-12775	CPD-4587
CPD-18152	CPD-16726	CPD-12776	CPD-4586
CPD-18148	CPD-16725	CPD-12774	CPD-10175
CPD-18147	CPD-16724	CPD-11751	CPD-7157
CPD-18145	CPD-16723	CPD-13381	CPD-10178
CPD-17342	CPD-16721	CPD-13952	CPD-4592
CPD-17343	CPD-16722	CPD-13380	STERIGMATOCYSTIN
CPD-17341	CPD-633	CPD-13389	6-DEMETHYLSTERIGMATOCYSTIN
CPD-17340	CPD-10176	CPD-13384	3-OH-BENZYL-ALCOHOL
CPD-17339	CPD-112	CPD-13383	CPD-10174
CPD-17337	CPD-16720	CPD-13382	CPD-10171
CPD-17336	CPD-10172	TYRAMINE	3-OH-BENZALDEHYDE
CPD-17335	CPD-637	CPD-10170	HYDRPHENYLAC-CPD
CPD-17334	CPD-402	CPD3O-4151	CPD-10169
CPD-17333	CPD-16600	CPD-13354	CPD-10168
CPD-17325	CPD-15209	CPD-10167	2-aminobenzoyl-CoA

CPD-17321	CPD-15213	CPD-12588	4-hydroxy-2-1H-quinolone
CPD-17317	CPD-14874	CPD-12836	CPD-10162
CPD-17318	CPD-14875	CPD-12838	CPD-10166
CPD-17316	CPD-14873	CPD-12708	CPD-10165
CPD-17074	CPD-7367	CPD-12709	CPD-10164
CPD-17014	CPD-14872	CPD-5923	CPD-10163
CPD-17019	GERANIOL	CPD-12706	CPD-10116
CPD-13896	CPD-8997	CPD-12707	CPD-10117
CPD-17018	CPD-4888	CPD-12710	CPD-9735
CPD-17017	geranial	CPD-9757	3-OXODODECANOATE
CPD-17016	CPD-7618	CPD-9699	2-UNDECANONE
CPD-17015	CPD-14319	CPD-16497	CPD-9700
CPD-16962	CPD-14321	CPD-7898	CPD-9727
CPD-16963	CPD-14320	CPD-14242	CPD-9734
CPD-16961	CPD-14322	CPD-7901	CPD-9736
CPD-16743	CPD-11890	INDOXYL	CPD-9737
CPD-16937	CPD-11786	CPD-10525	CPD-9738
CPD-16935	CPD-14323	CPD-10523	CPD-9739
CPD-16938	CPD-14318	INDICAN	CPD-9501
CPD-17455	CPD-14324	CPD-10230	CPD-9502
CPD-17053	VIOLACEIN	CPD-10231	ORCINOL-CPD
CPD-17453	CPD-13811	CPD-10229	CPD-14529
CPD-17052	CPD-13812	CPD-10228	CPD-7035
CPD-17051	CPD-6562	CPD-10227	CPD-14528
CPD-17050	CPD-6365	CPD-10226	CPD-9499
CPD-17049	CPD-6362	CPD-10225	CPD-9500
CPD-17048	CPD-6364	CPD-10224	CPD-9494
CPD-17047	CPD-6361	CPD-10216	CPD-16

Apéndice 4. Material suplementario.

Los programas escritos para este proyecto, así como los archivos de las redes se pueden descargar de:

http://eragene.com/tesis/

En caso de no estar disponibles favor de comunicarse al siguiente correo:

miguel.ramos@cinvestav.mx

Apéndice 5. Manual de usuario para el programa.

Continua a partir de la siguiente página.

2017

EraGene

MANUAL DE USUARIO VERSIÓN 1.0P

LABORATORIO DE INGENIERÍA BIOLÓGICA

eragene

Contenido

eragene

Índice de Imágenes

Interfaz gráfica

La interfaz gráfica de EraGene® está pensada para que el usuario pueda encontrar cualquier función a menos de 3 clics, es por esta razón que las funciones del programa están almacenadas en una barra de menús con varias pestañas (Ver **Imagen 1**), en las cuales se agrupan las funciones de acuerdo a alguna característica que compartan en común.

Las funciones disponibles del programa están representadas con iconos que representan la función de manera gráfica y además muestran un texto de ayuda cuando se pasa el puntero del ratón sobre ellos.

Cuando la función requiere de introducir o mostrar varios valores, de los botones se desplegarán menús emergentes con la información que se desea mostrar.

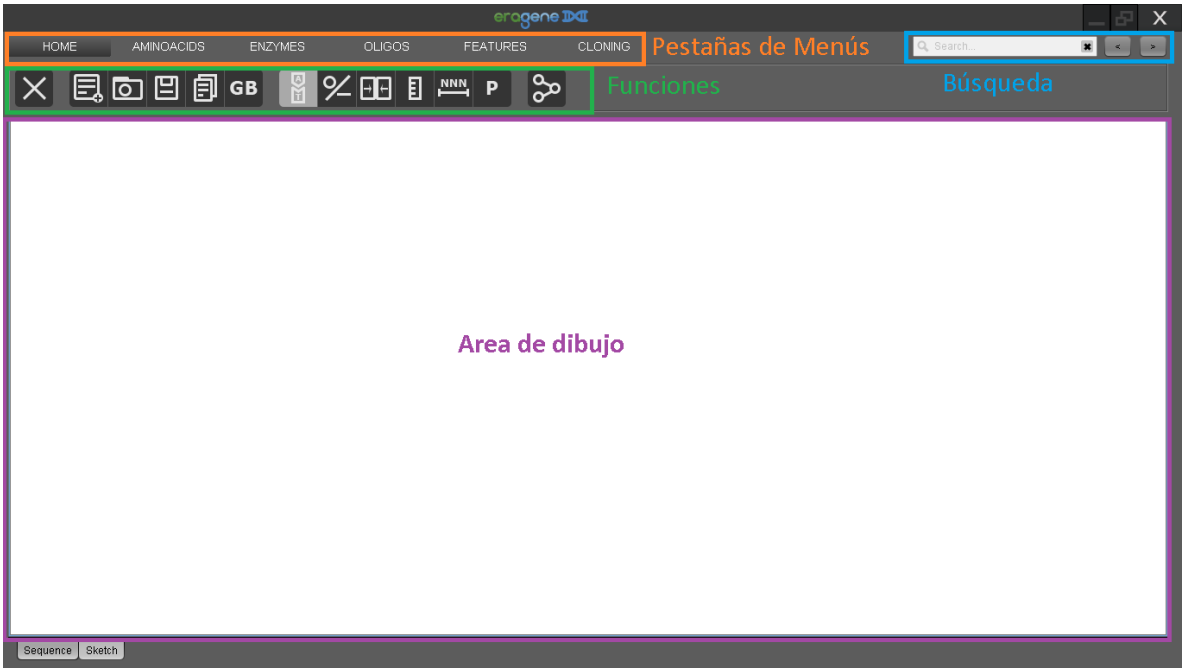


Imagen 1. Áreas que forman la interfaz gráfica del programa.

La primera impresión que tendrá de EraGene® es que utiliza un tema oscuro en el área de controles, esto se pensó así ya que en ocasiones el usuario tiene que pasar horas frente al computador analizando secuencias. Los colores oscuros son menos cansados para la vista por lo que evitará la fatiga en el usuario. En un futuro pensamos implementar un tema claro para los usuarios que así lo prefieran.

Menús

La barra de menú está inspirada en el menú Ribbon de Microsoft Office®, sin embargo, aquí se hace lo más angosta posible para que el área para visualizar las secuencias sea más grande y el usuario tenga que desplazar la imagen lo menos posible.

Imagen 2. Barra de menús.

Cómo se muestra en la **Imagen 2** hay varias pestañas en el menú, cada una de estas pestañas contiene funciones que comparten alguna característica en común, por ejemplo, la pestaña de aminoácidos contiene funciones que permiten visualizar los aminoácidos y buscar proteínas, la pestaña de enzimas tiene opciones para dibujar enzimas y simular cortes, etc.

Menús desplegables

Cuando una función del programa tiene opciones que puede seleccionar el usuario se despliega un menú del botón que tiene la función para que el usuario pueda seleccionarlas, y se cierra cuando el usuario da clic fuera del menú para que no estorbe a la vista (**Imagen 3**).

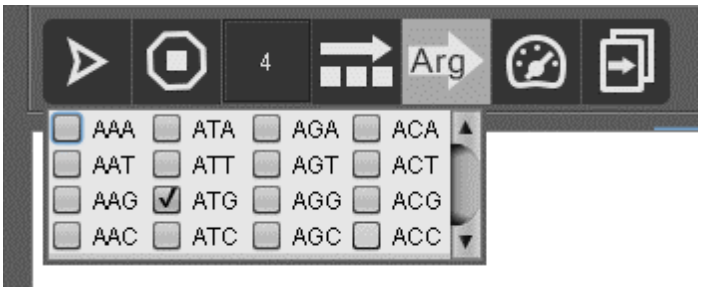


Imagen 3. Menús desplegables.

Herramienta de búsqueda

Imagen 4. Herramienta de búsqueda.

La herramienta de búsqueda a pesar de la sencilla apariencia que tiene permite al usuario realizar búsquedas de secuencias de ADN degeneradas (que utilizan nucleótidos no estándar) y no toma en cuenta si las letras están en mayúsculas o minúsculas, en lugar de tener un botón de búsqueda se puede activar la búsqueda dando un "*Enter*", además permite navegar entre los distintos sitios donde se encuentra la secuencia buscada con los botones que se muestran a la derecha en la **Imagen 4**.

En un futuro se espera añadir más funcionalidades como buscar enzimas por su nombre, anotaciones o aminoácidos.

Áreas de dibujo

Las áreas de dibujo (ver **Imagen 5**) son paneles donde se dibujan las secuencias de ADN o sus representaciones, junto con las anotaciones, estas áreas de dibujo no sólo se utilizan para dibujar secuencias, sino como el usuario descubrirá también son interactivas y permiten apuntar, seleccionar y modificar elementos junto con las funciones de los menús.

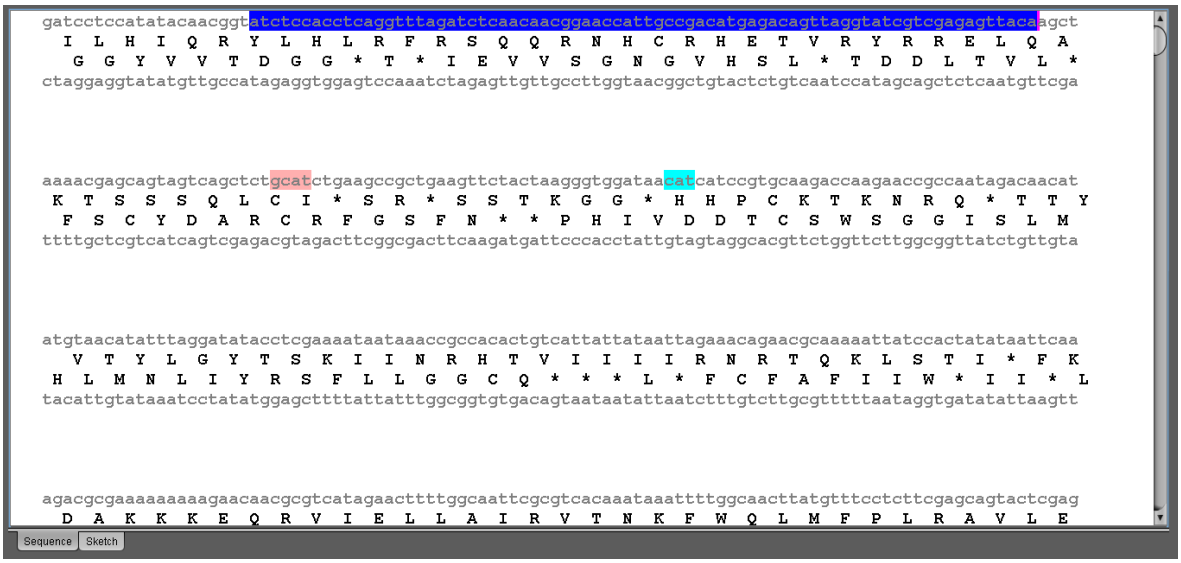


Imagen 5. Área de dibujo.

Introduciendo secuencias

Una de las características de EraGene® es que permite editar secuencias y para ello tiene varios métodos de introducirlas, desde hacerlo de manera automatizada hasta manual.

Crear secuencias desde cero

Para crear una nueva secuencia se puede introducir manualmente en el programa en la siguiente dirección HOME→New→Introducir nueva secuencia→Accept (Ver).

La ventana en la que se introduce la nueva secuencia nos permite seleccionar si esta secuencia es circular o no, además de que muestra información como el tamaño y el contenido de GC, también se revisa la secuencia introducida para encontrar secuencias no válidas de ADN.

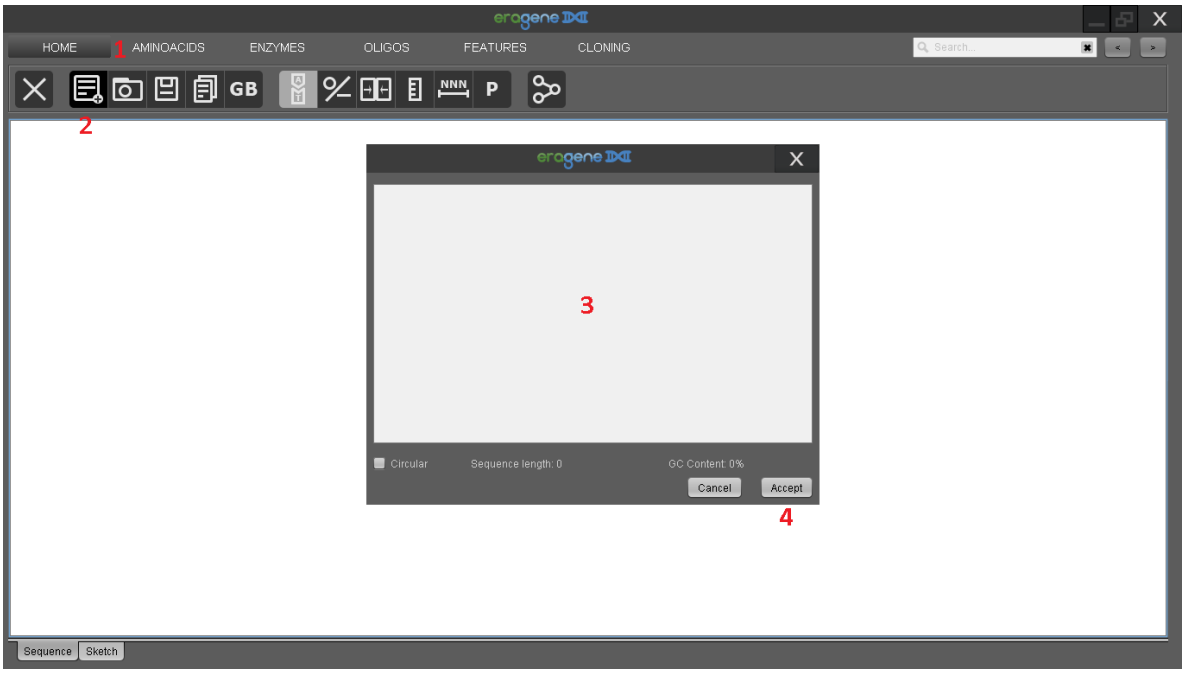


Imagen 6. Crear nueva secuencia.

Abrir archivos

EraGene® tiene soporte para archivos Fasta, soporte parcial para archivos GenBank y además utiliza su propio formato de archivos, para abrir un archivo es necesario ir a HOME→Open→Seleccionar archivo→Open (**Imagen 7**).

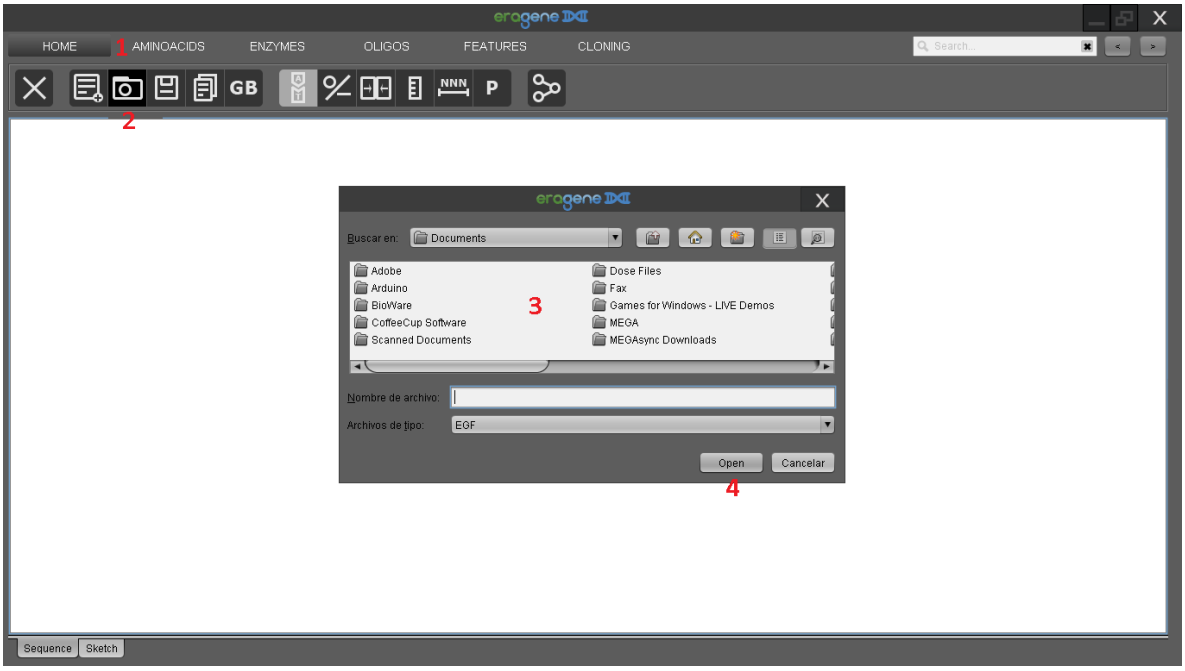


Imagen 7. Abrir un archivo.

Obtener secuencias de NCBI

Si trabaja con archivos GenBank de NCBI es posible que pueda descargar las secuencias utilizando su ID en HOME→GenBank→Introduzca su secuencia→Accept (Ver Imagen 8).

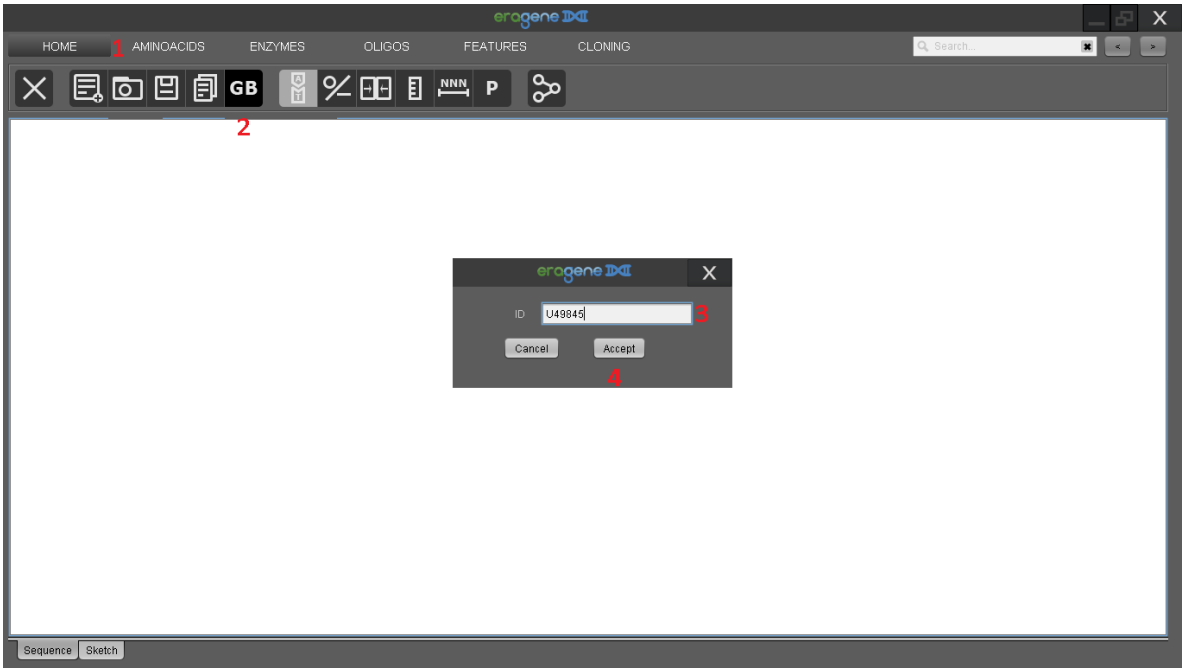


Imagen 8. Descargar archivo desde NCBI utilizando su ID.

Cómo esta función depende de NCBI puede ser que en ocasiones no esté disponible.

Reconstruyendo rutas metabólicas

Otra forma de introducir secuencias al programa es mediante la reconstrucción de rutas metabólicas, sin embargo, esté método es un poco más largo y requiere de mayor atención, por esa razón tiene su propio apartado.

Inicio

La reconstrucción de rutas metabólicas es una de las características más importantes de EraGene® ya que permite encontrar genes de miles de organismos tomados de bases de datos científicas, todo esto en una interfaz muy sencilla de utilizar que funciona en un proceso continuo. Básicamente se escribe un organismo en el cual se quiere producir un compuesto, se escribe el compuesto y se busca si el compuesto y organismo están presentes en la base de datos, para lo cual da una lista de posibles compuestos y organismos que puedan coincidir con la búsqueda y muestra información sobre ellos para aclarar que son los indicados, después se fija el número de reacciones máximas permitidas para alcanzar tal compuesto y se buscan las rutas posibles de síntesis (Ver Imagen 9).

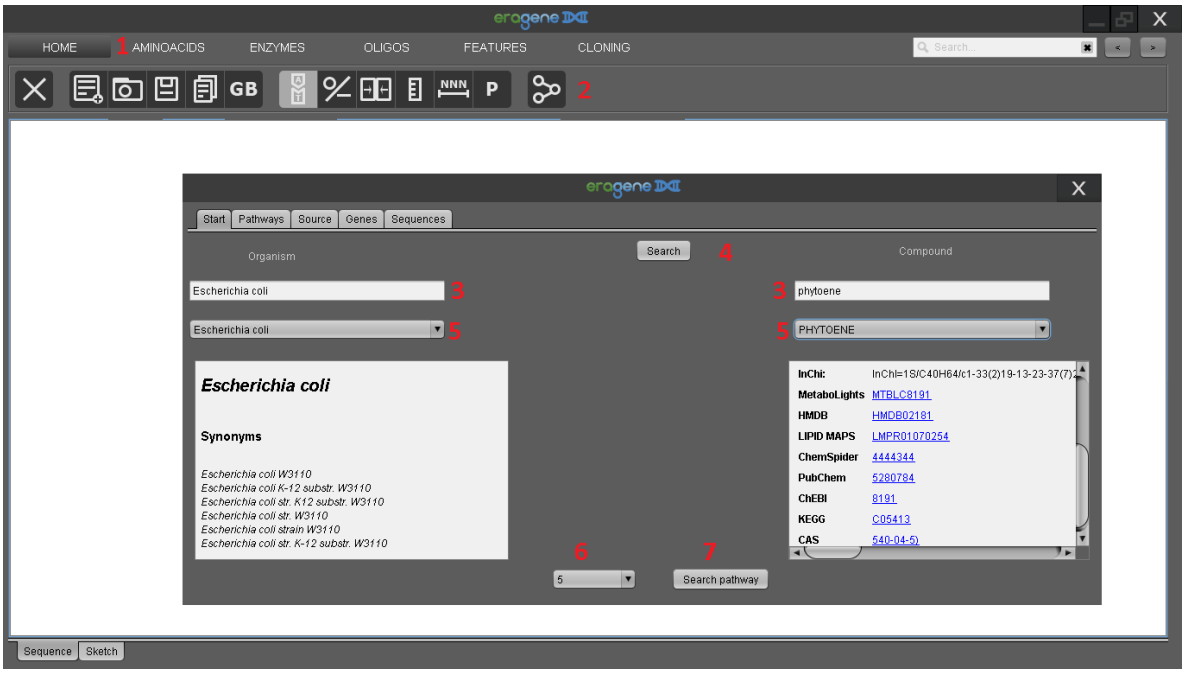


Imagen 9. Reconstruir rutas metabólicas (Start).

Rutas

Si se ha encontrado alguna ruta metabólica el programa la mostrará en la pestaña "*Pathways*", los compuestos deseados están en color rojo, las enzimas/reacciones en color blanco, los compuestos que no se producen normalmente en el organismo están en color gris, mientras que los que si se producen están en color cian. En ocasiones más de una familia de enzimas puede realizar una reacción por lo que al pasar el puntero del ratón sobre una enzima se mostrarán otras las familias que pueden llevar a cabo esa reacción. Cuando se ha seleccionado una ruta se puede dar clic en el botón que está sobre la ruta para continuar (Ver **Imagen 10**).

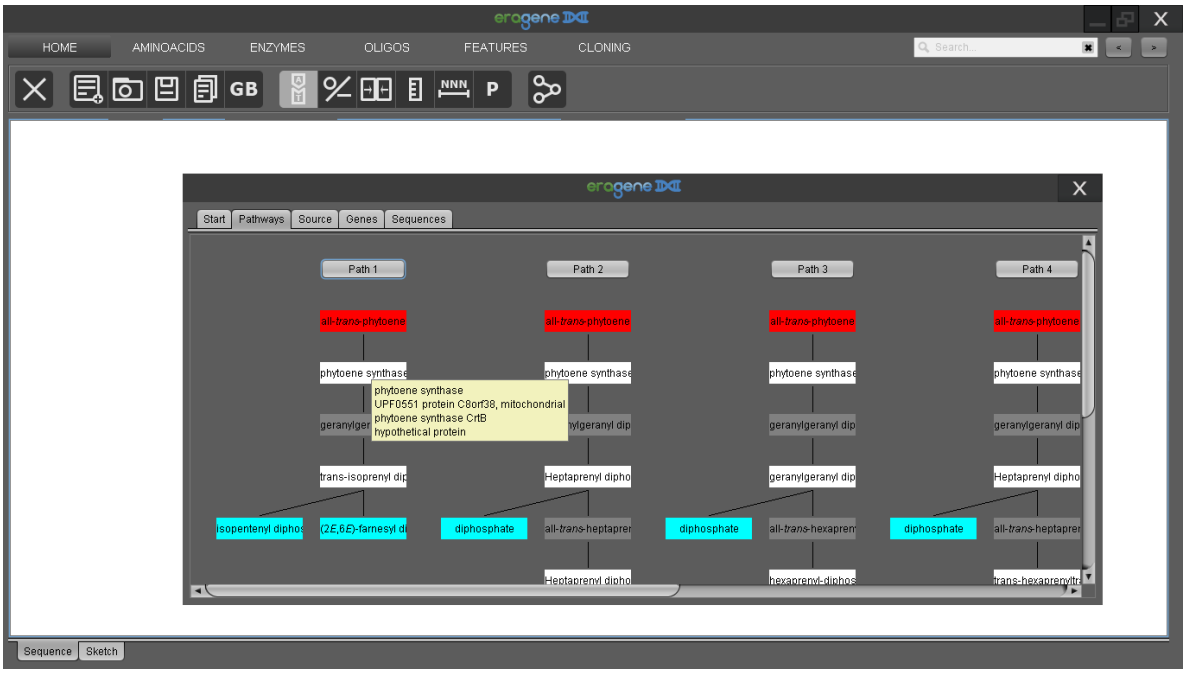


Imagen 10. Reconstrucción de rutas metabólicas (Pathways).

Genes

Una vez seleccionada la ruta para cada reacción el programa muestra las enzimas que pueden realizar estas reacciones. Estas enzimas se ordenan tomando en cuenta si han sido curadas manualmente, la cantidad de genes que requieren, la longitud de todos los genes, la cantidad de reacciones que llevan a cabo y la cantidad de nucleótidos que necesita. Para cada una de las enzimas se muestra el gen que la codifica, el nombre de la enzima y el organismo del que se obtiene, así

como la cantidad de nucleótidos que requiere. También se pueden consultar las enzimas en sus respectivas bases de datos haciendo clic en los enlaces.

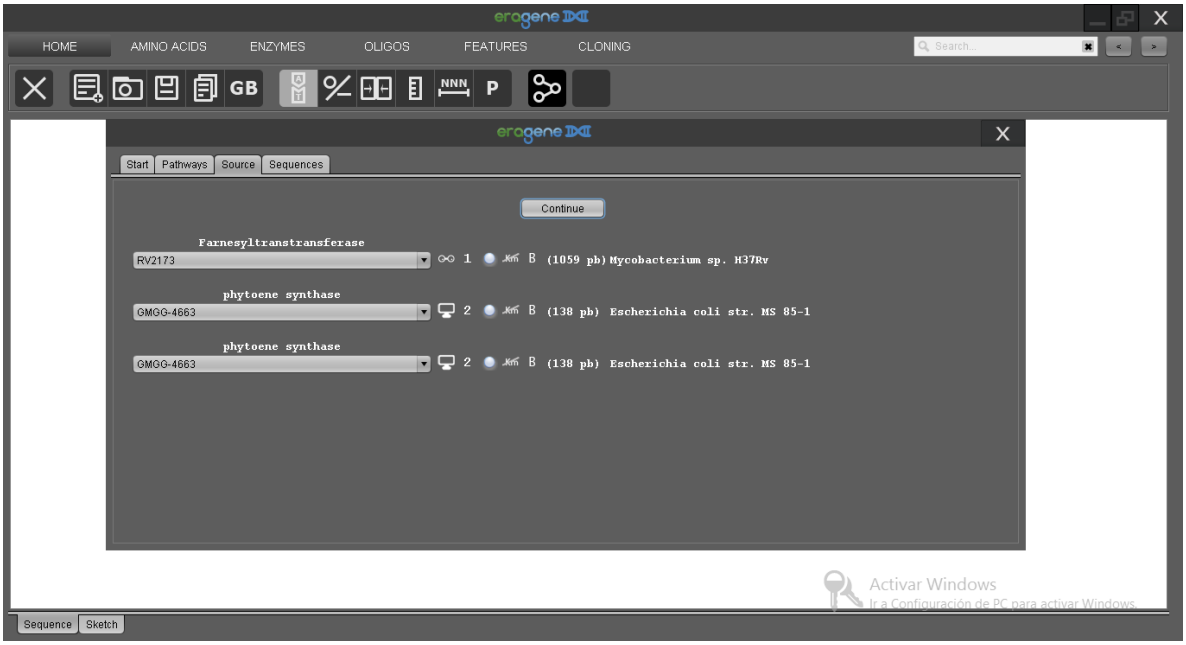


Imagen 11. Reconstrucción de rutas metabólicas (Source).

Secuencias

Finalmente, el programa encuentra los genes que codifican para las enzimas de las rutas metabólicas y da una opción para abrir cada secuencia con el editor de secuencias (Ver **Imagen 12**).

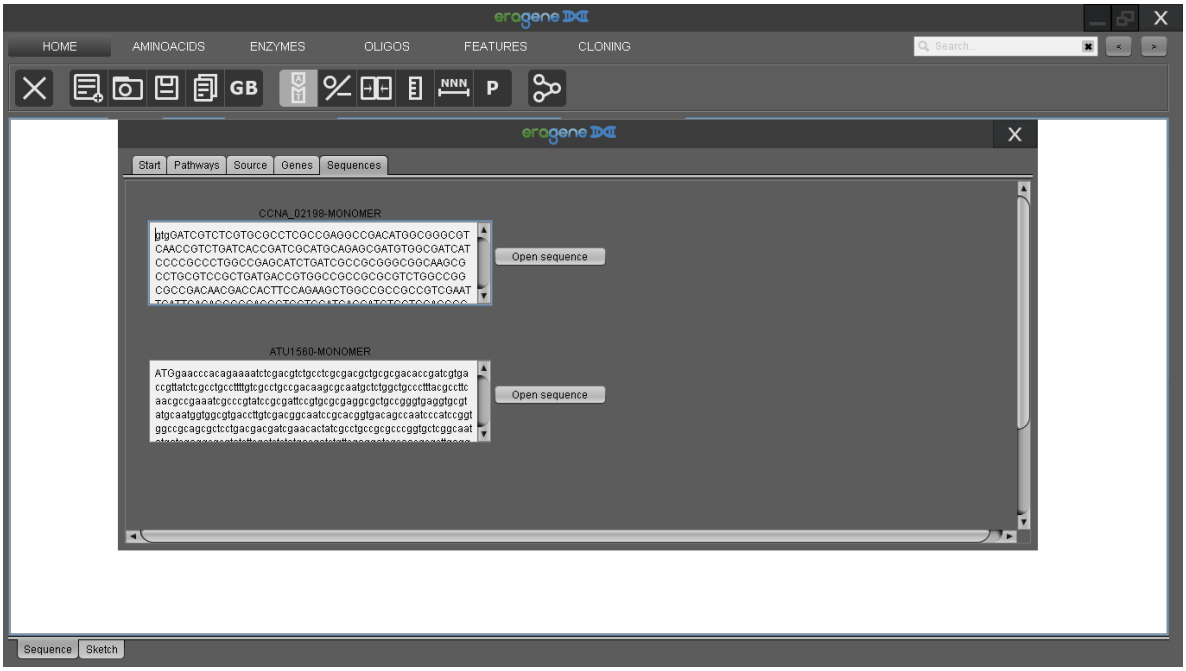


Imagen 12. Reconstrucción de rutas metabólicas (Sequences).

Visualización de secuencias

Las secuencias de ADN nos dicen poco a simple vista, por lo que hay que agregar anotaciones, para que sea más fácil para el usuario poder analizarlas, es por esta razón que EraGene® permite analizar las secuencias mediante visualizadores interactivos que permiten activar y desactivar el dibujado de distintos componentes.

Visualizador de secuencias

Las secuencias de ADN se pueden ver como texto interactivo, y mediante las funciones de los menús la cantidad de información que expresan las secuencias se puede modificar. Los dos tipos de visualizadores de secuencias permiten hacer operaciones básicas a la secuencia como copiar parte de las secuencias o realizar búsquedas en lo único que se diferencian es en que este visualizador permite ver las secuencias de ADN tal cual, mientras que en los otros sólo son representaciones gráficas.

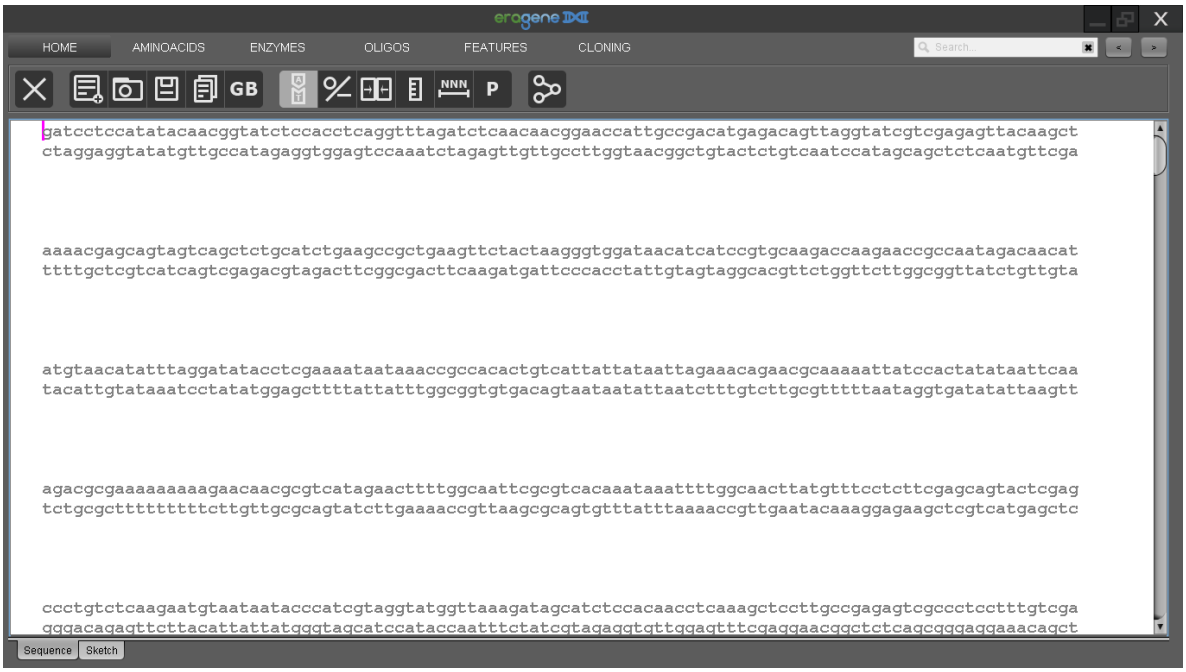


Imagen 13. Visualizador de secuencias.

Visualizador general

Este visualizador permite al usuario ver todas las anotaciones de la secuencia, así como las enzimas y oligos al mismo tiempo, por lo que es útil para determinar que anotación está detrás de otra o ubicar rápidamente un oligo, entre otras cosas. Este visualizador además de mostrar gráficamente las anotaciones también permite interactuar con ellas, así que se pueden apuntar o seleccionar para procesarlas (Ver Imagen 14).

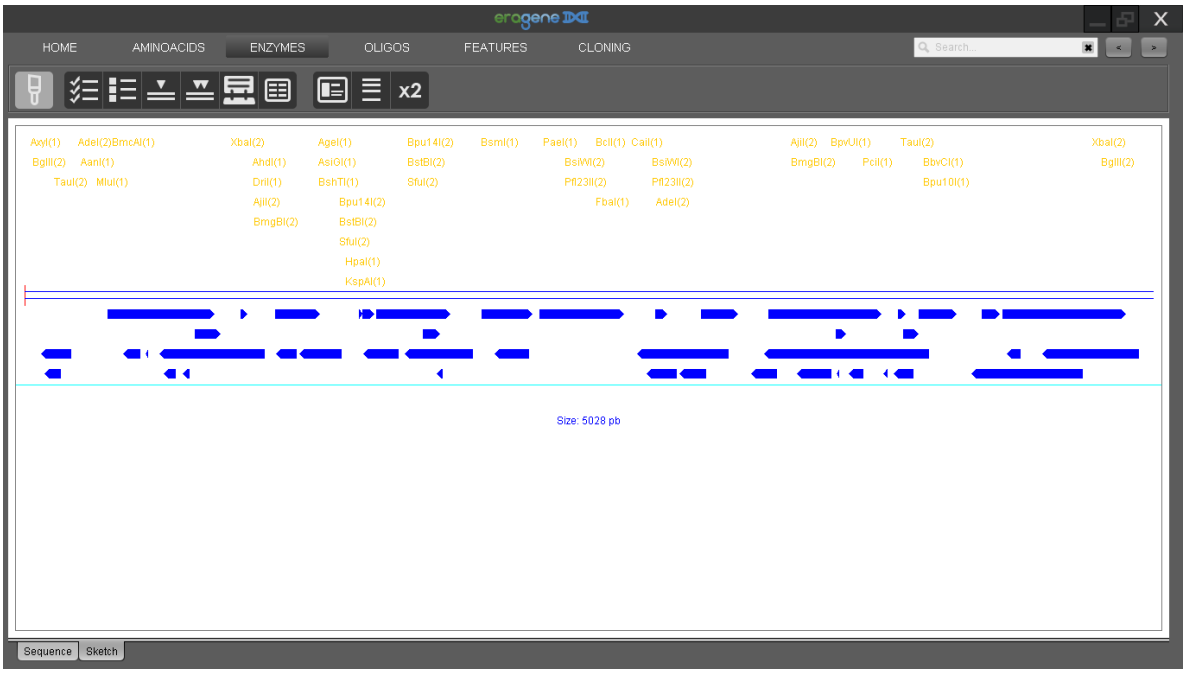


Imagen 14. Visualizador general.

Menú General

En el menú general se agrupan las funciones para el manejo de archivos, así como las funciones más simples de visualización de la secuencia.

Botones del menú general


Cierra la ventana a la que pertenece el botón.


Permite crear un nuevo archivo introduciendo una secuencia manualmente.


Muestra un dialogo para seleccionar archivos en formato fasta, Genbank o nativo.


Abre un cuadro de dialogo para guardar las secuencias en formato fasta, Genbank o nativo.


Muestra un menú desplegable que permite copiar las secuencias seleccionadas en distintas direcciones.

Descarga de NCBI un archivo Genbank utilizando su ID de registro.


Dibuja u oculta la secuencia complementaria de ADN.


Permite convertir una secuencia circular en lineal y viceversa.


Invierte la secuencia de la ventana, mostrando la secuencia complementaria de ADN como la principal.


Muestra una regla que permite identificar fácilmente cual es la posición de cada nucleótido.

Cambia la cantidad de nucleótidos que se muestran por línea en el visualizador de secuencias.

Abre un cuadro de dialogo donde se puede editar el estado de fosforilación de las secuencias.

Abre una nueva ventana que permite buscar rutas metabólicas y sus genes, se especifica con mayor profundidad en el tema Reconstruyendo rutas metabólicas.

Optimizar ORF

Para que un gen sea expresado de forma abundante en un nuevo organismo es necesario optimizar la secuencia del gen, para esto hay una función en el programa que le permite seleccionar un marco de lectura abierta y después abrir este marco de lectura en una nueva ventana para ser optimizado, la ventana que se crea contiene un editor con algunas herramientas, cómo por ejemplo crear una tabla de uso de codones que permite introducir el uso de codones que tiene el nuevo organismos, también se puede cambiar manualmente que codones se quieren cambiar en la secuencia o de forma automática cambiar todos los codones por los codones más usados o sólo cambiar los codones que están por debajo de un umbra. Una vez que se termina de editar la secuencia se puede guardar en un nuevo archivo (Ver Imagen 15).

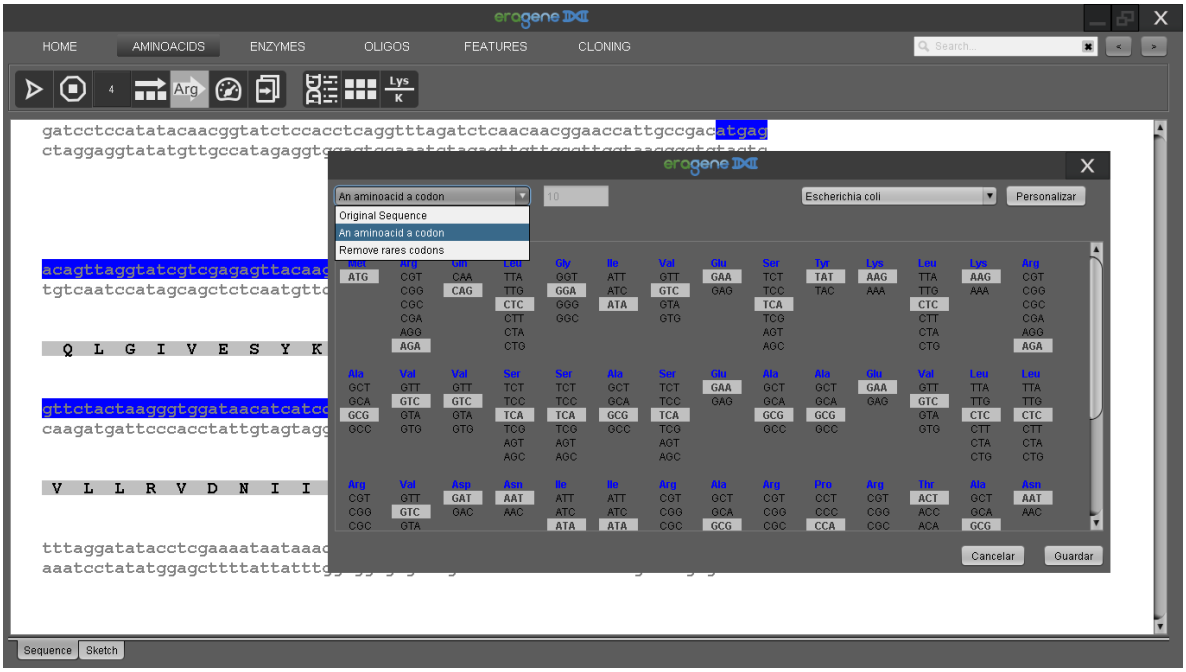


Imagen 15. Optimizador de secuencias.

Marcos de lectura abierta y aminoácidos

EraGene® tiene la posibilidad de mostrar las traducciones de las cadenas de ADN en todos sus marcos de lectura, además de que permite buscar Marcos de Lectura Abierta (ORF por sus siglas en inglés) que podrían ser genes y de una manera muy sencilla anotarlos. La búsqueda de marcos de

lectura es completamente personalizable y se pueden fijar todos los parámetros deseados para la búsqueda.

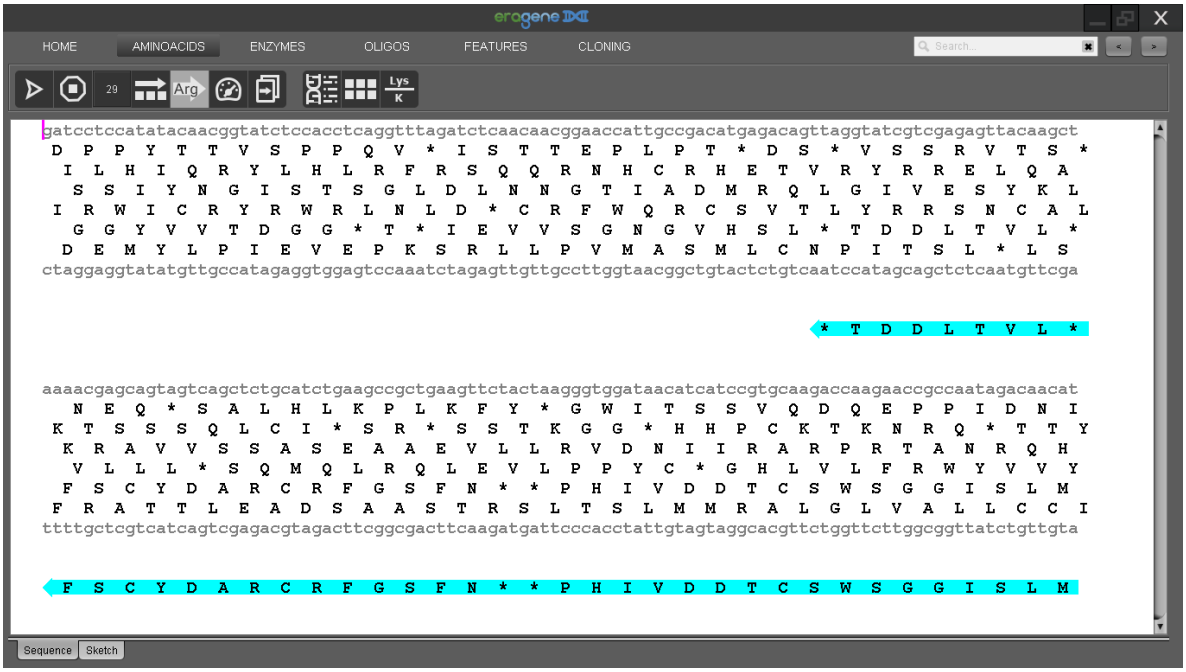


Imagen 16. Opciones de aminoácidos y marcos de lectura abierta.

Funciones de los botones en el menú de aminoácidos

Permite seleccionar los codones de inicio de los ORF a buscar.

Muestra la lista de codones de paro de los ORF a buscar.

Permite seleccionar el tamaño mínimo del ORF a buscar.

Selecciona los marcos de lectura para los cuales se buscarán los ORF.

Indica si se van a dibujar los aminoácidos sobre los ORF, en el Visualizador de Secuencias.

Permite optimizar el ORF que está seleccionado.

Permite copiar la secuencia de ADN o aminoácidos del ORF seleccionado.

Cambia el código genético con el cual se traducirán las secuencias.

Escoge los marcos de lectura en los cuales se mostrará la traducción de la secuencia de ADN.

Dibuja los aminoácidos en el código de 3 o 1 letra.

Enzimas

El menú de enzimas contiene un grupo de funciones relacionadas con la búsqueda de sitios de restricción, como son dibujar enzimas, simular geles de agarosa, comparar compatibilidad entre enzimas, entre otras.

Las enzimas dibujadas en los paneles pueden ser apuntadas y seleccionadas, además del nombre tienen la cantidad de veces que se han encontrado en toda la secuencia y permiten resaltar la secuencia de la enzima en la secuencia de ADN, así como las posiciones donde realizarán el corte las enzimas (Ver Imagen 17).

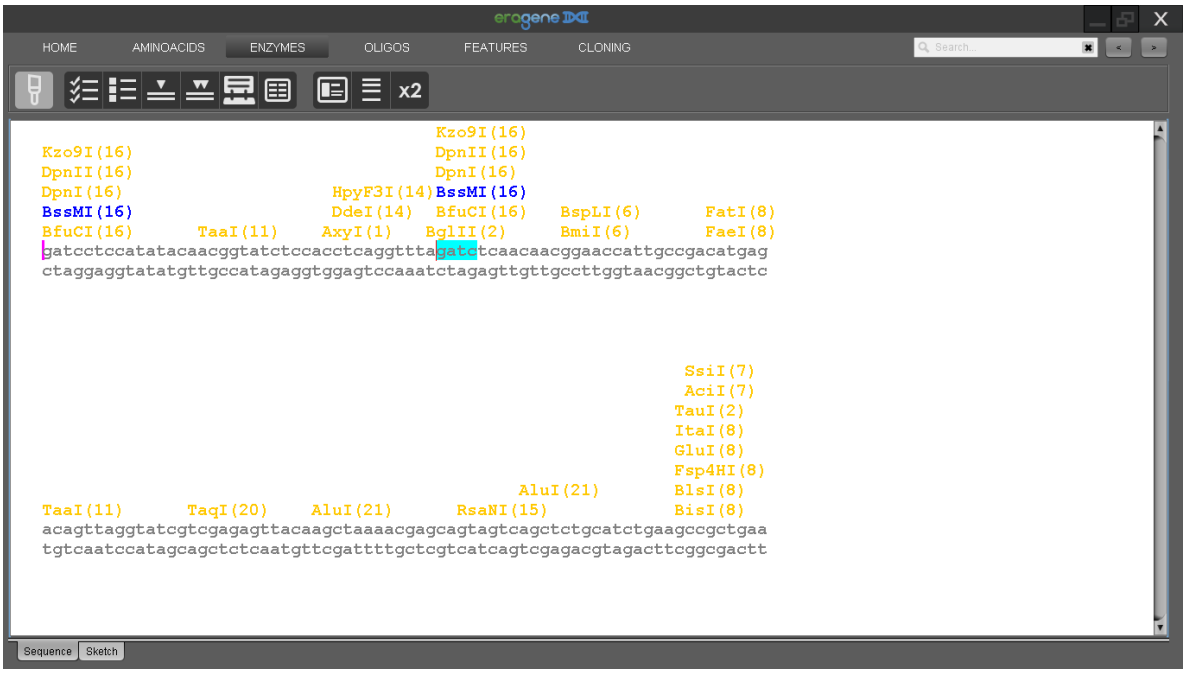


Imagen 17. Enzimas de restricción encontradas en la secuencia.

Botones del menú enzimas

Permite dibujar a no las enzimas encontradas

Añade todas las enzimas de restricción de la base de datos a la lista enzimas a buscar.

Quita todas las enzimas de la lista de enzimas a buscar.

Añade todas las enzimas que corten sólo una vez la secuencia.

Añade todas las enzimas que cortan dos veces la secuencia.

Muestra los perfiles disponibles para cargar enzimas de restricción.


Muestra la lista de todas las enzimas disponibles y las enzimas seleccionadas para buscar.


Muestra la información de todas las enzimas disponibles en la base de datos.


Simula un gel de agarosa con muestras de ADN.

Compara los parámetros de funcionamiento de dos enzimas, para realizar dobles digestiones.

Agregar y quitar enzimas

En el menú desplegable de enzimas seleccionadas hay dos listas, la izquierda contiene todas las enzimas disponibles para el programa, mientras que la lista derecha contiene las enzimas que serán dibujadas, el botón superior del panel permite agregar la enzima seleccionada en la lista izquierda a la lista derecha, mientras que el botón inferior permite borrar la lista seleccionada del panel derecho.

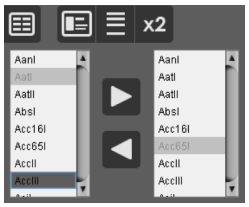
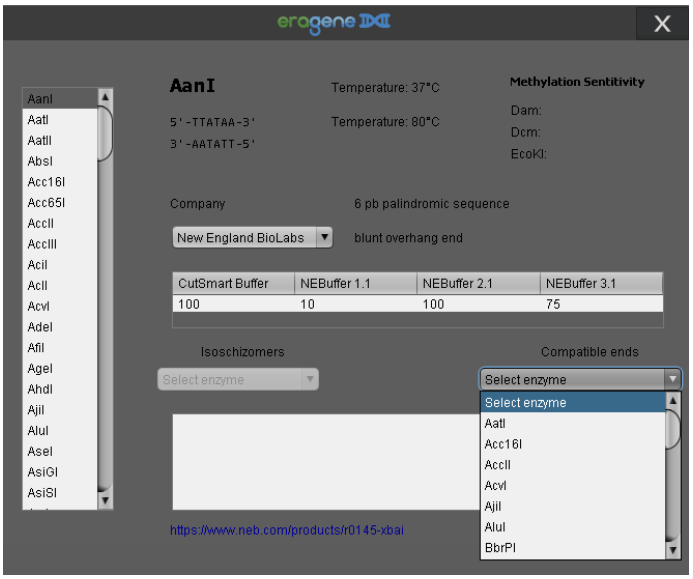


Imagen 18. Menú desplegable para editar enzimas a buscar.

Lista de enzimas disponibles



Hay una ventana que permite visualizar la información de las enzimas disponibles, en esta ventana hay una lista de todas las enzimas disponibles, cuando se selecciona una de estas enzimas se cargan sus parámetros en la ventana, se puede ver su secuencia, las compañías que las distribuyen, los buffers que utilizan, la temperatura de funcionamiento, el tipo de metilación al cual son sensibles, además calcula cuales enzimas tienen el mismo sitio de reconocimiento o con cuales son compatibles (Ver Imagen 19). Toda esta información es obtenida de un archivo.

Imagen 19. Enzimas disponibles.

Simulación de electroforesis en gel para ADN

Se pueden simular geles de agarosa con secuencias de archivos, además se pueden cargar marcadores de peso molecular para comparar los pesos de las secuencias. Es posible simular el corte simultaneo de hasta 4 enzimas, además de que crea una lista de bandas que se generarán con los cortes en todo el gel. La imagen es interactiva y se pueden seleccionar los carriles para añadir secuencias o se pueden resaltar las bandas de acuerdo a su tamaño.

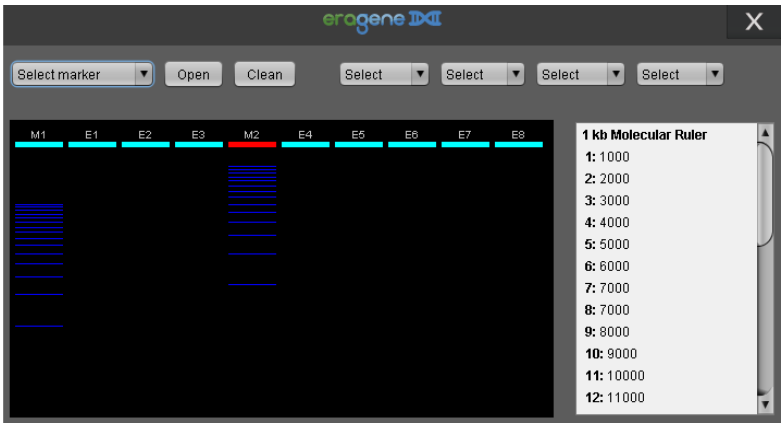


Imagen 20. Simulación de geles de agarosa.

Comparación de parámetros de incubación para enzimas

En el laboratorio a menudo es necesario cortar ADN con dos enzimas simultáneamente, por esta razón hay una ventana que permite comparar los parámetros de dos enzimas, para seleccionar cuales son los mejores parámetros para trabajar.

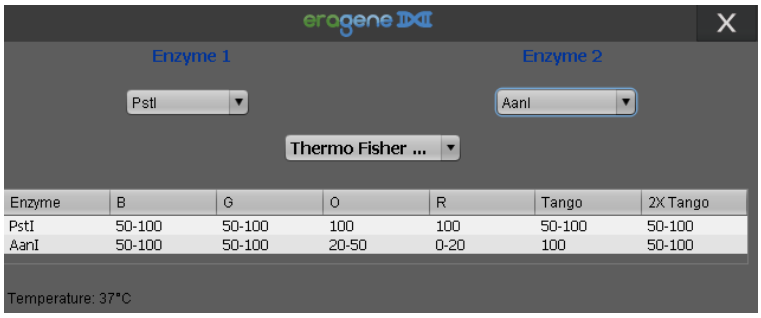


Imagen 21. Comparación de enzimas para doble digestión.

Oligos

El programa permite anotar oligos sobre las secuencias de ADN, tanto para visualizarlos como para simular reacciones de PCR, además calculo parámetros importantes para el diseño de oligos.

Descripción de los botones del menú Oligos

Selecciona si se deben pintar los oligos del archivo o no.

Abre una ventana en la cual se pueden añadir más oligos.

Elimina el oligo seleccionado.

Añadir nuevos oligos

Para añadir un oligo es necesario seleccionar el botón para añadir oligos, cuando hay secuencias seleccionadas se desplegará un menú con la opción de copiar la secuencia principal o

complementaria para empezar a diseñar el oligo, si no hay selección de texto se abre directamente una nueva ventana para editar el oligo, como se muestra en la Imagen 22.

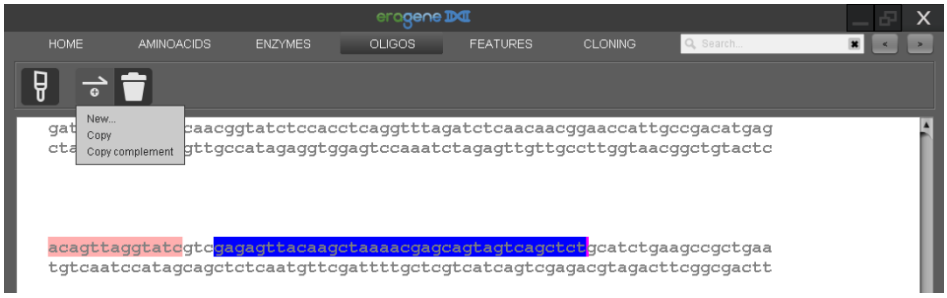


Imagen 22. Añadir nuevo oligo.

La ventana de edición de oligos permite editar los oligos, cada uno de ellos tendrá una secuencia y nombre, para los cuales se calcularán de manera automática parámetros termodinámicos y se buscará la secuencia en el archivo abierto para determinar que no haya productos secundarios u otros problemas que impidan realizar PCRs.

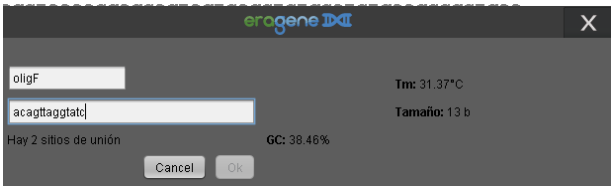


Imagen 23. Ventana de edición de oligos.

Anotaciones

Las secuencias de ADN tienen regiones que representan genes, sitios de reconocimiento, promotores, etc., y es necesario resaltarlas para que sean fáciles de identificar por los usuarios, por esta razón existen las anotaciones, en el programa podrás añadir fácilmente anotaciones y estas son parcialmente compatibles con el formato de GenBank por lo que podrá guardarlas y recuperarlas de archivos.

Descripción de los botones del menú anotaciones

Indica si se van a dibujar o no las anotaciones.

Permite añadir una nueva anotación a partir de la selección de una secuencia.

Elimina la anotación seleccionada.

Oculta la anotación que esta siendo seleccionada.

Vuelve visibles todas las anotaciones que han sido ocultadas.

Cambia el color de la anotación que esta seleccionada.

Añadir nuevas anotaciones

Para añadir nuevas anotaciones es necesario realizar una selección sobre la secuencia de ADN y añadir la anotación, después se abrirá una ventana que permite cambiar la información de la anotación, se puede especificar qué tipo de anotación es por ejemplo promotor, gen, entrón, etc. Además, se puede especificar en donde empieza y termina la anotación, el color y si esta sobre la cadena complementaria.

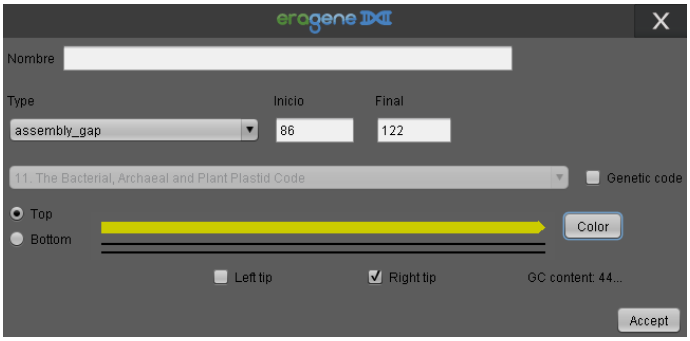


Imagen 24. Ventana para añadir anotaciones.

Clonación

Para que el usuario pueda planear los métodos de ensamble el programa tiene una función para digestiones y ligaciones de plásmidos, además de que se pueden también simular PCRs.

Funciones de los botones del menú clonación

Cuando hay dos oligos seleccionados permite crear un nuevo archivo a partir del producto de la PCR.

Abre una ventana que permite simular una clonación utilizando enzimas de restricción.

PCR

Para simular PCRs es necesario primer añadir oligos a la secuencia, una vez añadidos los oligos, se deben seleccionar; el primer oligo se selecciona con un clic, mientras que el segundo se debe seleccionar con un clic mientras se presiona simultáneamente la tecla "*Shift*". Después se debe presionar el botón de PCR para validar si los oligos son válidos y si lo son se creará un nuevo archivo con la secuencia de ADN amplificada, ver Imagen 25.

Imagen 25. Simular PCR.

Clonación por restricción

En el programa se pueden simular clonaciones con enzimas de restricción, utilizando una secuencia que sirve como inserto y otra como vector. En la ventana de clonación hay tres paneles que contienen los archivos a dibujar, cuando uno de estos paneles esta seleccionado se activaran los botones que están en la parte superior de la ventana para ese panel, para simular la clonación se debe n seleccionar dos enzimas en cada archivo para esto se selecciona la primera enzima con un clic y la segunda con un clic mientras se presiona la tecla "*Shift*", si los extremos generados por las enzimas son compatibles se generará un nuevo archivo que puede ser visualizado en el panel de resultados.

Imagen 26. Simular clonación usando enzimas de restricción.