

Instituto Tecnológico de Costa Rica  
Escuela de Ingeniería en Computación  
Programa de Maestría en Computación



# **Diseño de una Herramienta para la Asistencia en el Ensamblaje Sintético de la Bacteria E. Coli en la Producción de Biocombustibles**

Tesis sometida a consideración del Departamento de  
Computación, para optar por el grado de Magister Scientiae  
en Computación con énfasis en Ciencias de la Computación

Estudiante:  
**Laura M. Vásquez Rodríguez**

Profesor Asesor:  
Dr. Allan Orozco

Cartago, Costa Rica  
Junio, 2014

## RESUMEN

Por el impacto que representa la contaminación ambiental, actualmente resulta de gran valor la producción de biocombustibles de calidad así como el aprovechamiento de desechos orgánicos que normalmente no tienen ninguna utilidad.

A través de la Biología Sintética, que es la ciencia que diseña y construye nuevos sistemas biológicos, es posible el mejoramiento de organismos existentes para procesar desechos contaminantes que tradicionalmente no se utilizan en el dominio de los biocombustibles, tales como el suero de la leche.

Las rutas metabólicas son un conjunto de reacciones químicas catalizadas por enzimas que permiten llevar a cabo funciones esenciales dentro de las células. A través del rediseño de estas rutas, es posible crear conexiones que permiten la implementación de nuevas funciones que normalmente no se encuentran presentes en la naturaleza. Desde un punto de vista computacional, es posible la traducción de rutas metabólicas, las cuales se encuentran almacenadas en repositorios como la Enciclopedia de Genes y Genomas de Kyoto (KEGG), a estructuras de datos básicas, tales como los grafos. Estas transformaciones habilitan la aplicación de los algoritmos ya estudiados en el área, lo que permite contribuir a la optimización de los análisis necesarios para alcanzar el producto final.

El presente trabajo tiene como objetivo el diseño de una herramienta que permita la transformación de rutas metabólicas y el desarrollo de algoritmos de búsqueda que establezcan rutas relevantes entre compuestos relativos a la producción de biocombustibles. Como resultado, se creará un catálogo de componentes genéticos, o biobricks, encontrados a partir del análisis de rutas en particular, para el ensamblaje sintético en la bacteria *E. Coli*. Los ensamblajes se caracterizarán estructuralmente y funcionalmente según las piezas utilizadas. Finalmente, se visualizarán las nuevas construcciones con el fin de mostrar y soportar el proceso de análisis, asistiendo a los profesionales del área que llevan a cabo este trabajo.

**Palabras Claves:** *Biología Sintética, grafos, rutas metabólicas, bioinformática, KEGG, biocombustibles, proteínas, biobricks.*

## **ABSTRACT**

Due to the impact of environmental pollution, it is valuable to produce high quality biofuels and to leverage organic waste that normally would have no use.

Through Synthetic Biology, the science of designing and building new biological systems, it is possible to improve existing organisms to process waste that is not traditionally used for biofuels production, such as butter milk.

Metabolic pathways are a set of chemical reactions catalyzed by enzymes that carry out essential functions within cells. Through the redesign of these pathways, it is possible to create connections that allow for the implementation of new functions that normally are not present in nature. From a computational point of view, metabolic pathways, which can be found in data sources as the Encyclopedia of Genes and Genomes (KEGG), can be converted to basic data structures, such as graphs.

These transformations enable the application of well known algorithms, which allows for the optimization of the analyses required to achieve the assembly of new organisms.

The present work aims to design a tool that allows for the transformation of metabolic pathways and the development of path finding algorithms that establish relevant links between compounds that are essential to the production of biofuels.

As a result, a catalog of biobricks will be created from the analysis of a subset of paths. This will allow for the synthetic assembly of the E. Coli bacteria. The assembly's structure and functions will be characterized according to the pieces used. Finally, new constructions will be visualized with the goal of demonstrating and supporting the analysis processes, thus, assisting people that work in the field of Synthetic Biology.

**Keywords:** *Synthetic Biology, graphs, metabolic pathways, bioinformatics, KEGG, biofuels, proteins.*

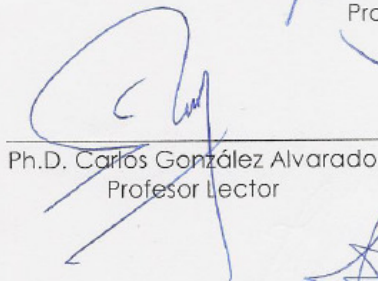
APROBACIÓN DE LA TESIS

**“Diseño de una Herramienta para la Asistencia en el  
Ensamblaje Sintético de la Bacteria E. Coli en  
la Producción de Biocombustibles”**

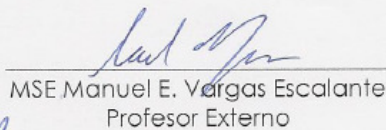
TRIBUNAL EXAMINADOR



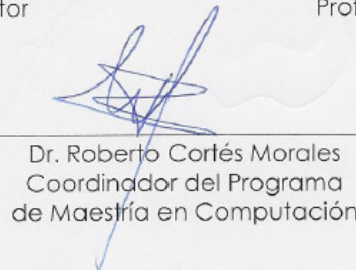
Ph.D. Allan Orozco Solano  
Profesor Asesor



Ph.D. Carlos González Alvarado  
Profesor Lector



MSE Manuel E. Vargas Escalante  
Profesor Externo



Dr. Roberto Cortés Morales  
Coordinador del Programa  
de Maestría en Computación

Junio, 2014

*A mis padres, a mi hermano Orlando y a Char,*

*Gracias por el apoyo incondicional y la motivación que me han dado a lo largo de estos años, porque de una forma u otra, cada uno de ustedes me enseñó que cuando soñamos y queremos lograr una meta en nuestras vidas, no hay imposible.*

## **AGRADECIMIENTOS**

Al Dr. Allan Orozco, quien como profesor asesor, me brindó la oportunidad de realizar el presente proyecto y sobre todo, por su apoyo, guía y motivación para cumplir las metas propuestas.

Al Ing. Ricardo Alvarado, por su importante colaboración en el diseño de la herramienta y el prototipo, así como su apoyo durante todo el proyecto.

A los Ing. David García, Ing. Abad Rodríguez y al estudiante Silver Ceballos, por compartir sus proyectos y sus conocimientos en el área de la Biología Sintética. Sus aportes fueron esenciales para el planteamiento de la presente investigación.

Al M.Eng Manuel Vargas Escalante, M.Sc. Diego Rivera-Gutiérrez, Ing. Carlos Álvarez, MBA. Mayra Rodríguez y M.Sc. Shannon Cummings por su colaboración en la revisión de la tesis y su retroalimentación, las cuales fueron de gran valor para el presente trabajo.

Al Dr. Carlos González por su guía en el planteamiento y revisión del anteproyecto de la presente investigación.

Al Consejo Nacional para Investigaciones Científicas y Tecnológicas (CONICIT), quién brindó un apoyo financiero significativo para el desarrollo de la presente maestría.

A la empresa Componentes Intel de Costa Rica, por brindarnos la oportunidad de realizar la maestría en alianza con el TEC y por su apoyo económico complementario para el desarrollo de la presente maestría.

A María Catalina Rosales del Centro Nacional de Innovaciones Biotecnológicas (CENIBiot), al M.Sc. Juan Carlos Saborío-Morales del Laboratorio Nacional de Computación Avanzada (CNCA), María Mora del INBIO y a Mayela Guzmán del Departamento de Admisión y Registro del TEC, por su colaboración al planteamiento de los antecedentes de este proyecto.

*Cree que puedes y estás a mitad de camino*

Believe you can and you're halfway there

-Theodore Roosevelt

# ÍNDICE

<b>CAPITULO 1. DEFINICIÓN DEL PROYECTO .....</b>	<b>15</b>
1.1. PLANTEAMIENTO DEL PROBLEMA .....	15
1.2. OBJETIVOS.....	18
1.2.1. <i>Objetivo General</i> .....	18
1.2.2. <i>Objetivos Específicos</i> .....	18
1.3. JUSTIFICACIÓN.....	19
1.4. HIPÓTESIS .....	21
<b>CAPITULO 2. FUNDAMENTACIÓN TEÓRICA .....</b>	<b>22</b>
2.1. LA BIOINFORMÁTICA EN COSTA RICA .....	22
2.2. LA BIOLOGÍA SINTÉTICA Y LOS BIOCOMBUSTIBLES .....	26
2.3. LA BIOLOGÍA SINTÉTICA EN COSTA RICA.....	32
2.4. HERRAMIENTAS DE DISEÑO .....	34
<b>CAPITULO 3. METODOLOGÍA.....</b>	<b>37</b>
3.1. DESCRIPCIÓN DEL PROCESO.....	37
3.2. DEFINICIÓN DE ETAPAS.....	39
3.2.1. <i>Etapa 1: Análisis de rutas metabólicas</i> .....	40
3.2.2. <i>Etapa 2: Selección de proteínas</i> .....	42
3.2.3. <i>Etapa 3: Selección de genes</i> .....	43
3.2.4. <i>Etapa 4: La creación de un catálogo y sus ensamblajes</i> .....	46
3.2.5. <i>Etapa 5: Visualización de ensamblajes</i> .....	50
3.3. DELIMITACIÓN DE ENTRADAS.....	50
<b>CAPITULO 4. IMPLEMENTACIÓN DEL PROTOTIPO .....</b>	<b>58</b>
4.1. DISEÑO DEL MODELO COMPUTACIONAL.....	58



4.2. EXTRACCIÓN DE DATOS .....	63
4.3. RUTAS METABÓLICAS .....	66
4.3.1. <i>Transformación de entradas</i> .....	66
4.3.2. <i>Construcción de rutas a través de grafos</i> .....	67
4.3.3. <i>Algoritmos de búsqueda</i> .....	76
4.4. SELECCIÓN DE PROTEÍNAS.....	87
4.5. SELECCIÓN DE GENES.....	92
4.6. CONSTRUCCIÓN DEL CATÁLOGO DE BIOBRICKS.....	98
4.7. VISUALIZACIÓN DE LOS DATOS .....	105
4.8. ARCHIVOS DE SALIDA .....	106
4.9. TECNOLOGÍAS UTILIZADAS.....	108
<b>CAPITULO 5. ANÁLISIS DE RESULTADOS .....</b>	<b>109</b>
5.1. RUTAS METABÓLICAS .....	110
5.2. PROCESO DE CREACIÓN DEL CATÁLOGO Y LOS ENSAMBLAJES .....	116
5.3. ESCALABILIDAD DE LA SOLUCIÓN .....	123
5.4. LIMITACIONES .....	127
5.4.1. LIMITACIONES DEL PROTOTIPO.....	127
5.4.2. LIMITACIONES DE LOS DATOS DE ENTRADA .....	131
<b>CAPITULO 6. CONCLUSIONES Y RECOMENDACIONES .....</b>	<b>136</b>
6.1. CONCLUSIONES.....	136
6.2. RECOMENDACIONES .....	140
<b>BIBLIOGRAFÍA .....</b>	<b>143</b>
<b>APÉNDICES.....</b>	<b>149</b>
APÉNDICE 1. MAPA DE LA GLUCÓLISIS EN LA BACTERIA E. COLI .....	149
APÉNDICE 2. MAPA DE LA BIOSÍNTESIS DE LA ESTREPTOMICINA EN LA BACTERIA E. COLI.....	150

APÉNDICE 3. MAPA DEL METABOLISMO DEL CARBONO EN LA BACTERIA E. COLI .....	151
APÉNDICE 4. MAPA DEL METABOLISMO DE LA GALACTOSA EN LA BACTERIA E. COLI .....	152
APÉNDICE 5. MAPA DE LA RUTA DE LA PENTOSA FOSFATO EN LA BACTERIA E. COLI .....	153
APÉNDICE 6. ARCHIVO EN FORMATO KGML DE LA GLUCÓLISIS EN BACTERIA E. COLI .....	154
APÉNDICE 7. DESCARGA DE ARCHIVOS KGML.....	155
APÉNDICE 8. LISTA DE RUTAS METABÓLICAS PARA LA LACTOSA Y PIRUVATO. ....	156
APÉNDICE 9. CATEGORÍAS Y BIOBRICKS SELECCIONADOS (iGEM FOUNDATION, 2014).....	159
APÉNDICE 10. GLOSARIO .....	161

## ÍNDICE DE FIGURAS

FIGURA 1. NÚMERO DE GRADUADOS EN LAS CARRERAS DE INGENIERÍA EN COMPUTACIÓN Y EN BIOTECNOLOGÍA DEL TEC DESDE AÑO 2000 AL 2013 .....	24
FIGURA 2. BIOLOGÍA SINTÉTICA (HEINEMANN & PANKE, 2006) .....	27
FIGURA 3. PROCESO DE LA PRODUCCIÓN DE BIOETANOL (CHIN, 2008) .....	29
FIGURA 4. LOS ORGANISMOS Y SUS PROCESOS.....	37
FIGURA 5. DISEÑO DE NUEVAS RUTAS METABÓLICAS .....	38
FIGURA 6. REGIÓN CODIFICANTE PARA EL GEN GMP1 (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 1994) .....	44
FIGURA 7. PREFIJO Y SUFIJO PARA UNA SECUENCIA CDS.....	45
FIGURA 8: PREFIJO Y SUFIJO PARA UNA SECUENCIA NO CDS.....	46
FIGURA 9. BIOBRICK ESTÁNDAR CON SU PREFIJO, SUFIJO Y SITIOS DE CORTE (INTERNATIONAL GENETICALLY ENGINEERED MACHINE FOUNDATION, 2003) .....	47
FIGURA 10. ENSAMBLAJE ESTÁNDAR DE BIOBRICKS .....	48
FIGURA 11. MODELO COMPUTACIONAL .....	59
FIGURA 12. ARQUITECTURA DE LA APLICACIÓN.....	62
FIGURA 13. URL ESTÁNDAR PARA LA DESCARGA DE MAPAS DE KEGG .....	65
FIGURA 14. PROCESO DE TRANSFORMACIÓN DE DATOS.....	68
FIGURA 15. GRAFOS ANIDADOS DEBIDO A LA AGRUPACIÓN DE ENZIMAS .....	69
FIGURA 16. GRAFOS ANIDADOS TRANSFORMADOS EN CAMINOS INDEPENDIENTES.....	70
FIGURA 17. UNIÓN DE GRAFOS INDEPENDIENTES EN UNA SOLA ESTRUCTURA.....	72
FIGURA 18. CARGA DE MAPAS, BIOBRICKS Y PLÁSMIDOS. ....	73
FIGURA 19. CARGA DE RUTAS METABÓLICAS EN FORMATO GRAPHML .....	74
FIGURA 20. UNIÓN DE RUTAS METABÓLICAS EN UN SÓLO GRAFO .....	74
FIGURA 21. EXTRACCIÓN DE DATOS DE KEGG POR MEDIO DEL API REST.....	75
FIGURA 22. BÚSQUEDA POR ANCHURA .....	78
FIGURA 23. PSEUDOCÓDIGO DE BÚSQUEDA POR ANCHURA PARA TODOS LOS CAMINOS .....	80

FIGURA 24. VISUALIZACIÓN DEL CAMINO MÁS CORTO ENTRE LA LACTOSA Y EL PIRUVATO .....	83
FIGURA 25. SELECCIÓN DE ENTRADAS PARA LA BÚSQUDA DE CAMINOS EN MODO PERSONALIZADO .....	85
FIGURA 26. VISUALIZACIÓN DEL CAMINO INCLUYENDO REACCIONES Y PROTEÍNAS.....	86
FIGURA 27. SELECCIÓN DE LA RUTA METABÓLICA DE INTERÉS .....	88
FIGURA 28. PROCESO DE EXTRACCIÓN DE INFORMACIÓN DE LAS PROTEÍNAS.....	92
FIGURA 29. EBOT HOME PAGE (SAYERS, 2010).....	94
FIGURA 30. GENERACIÓN DEL SCRIPT PARA LA DESCARGA DEL CDS POR MEDIO DE EBOT (SAYERS, 2010) .....	96
FIGURA 31. EXTRACCIÓN DE LA REGIÓN CODIFICANTE DEL GEN .....	97
FIGURA 32. BÚSQUDA Y CARGA DE GENES A PARTIR DE REFERENCIAS .....	97
FIGURA 33. PIEZAS DEL CATÁLOGO INGRESADAS POR EL USUARIO .....	99
FIGURA 34. EXCLUSIÓN DE REGIONES CODIFICANTES .....	100
FIGURA 35. CATÁLOGO PARA ENSAMBLAJES .....	101
FIGURA 36. MODALIDADES DE ENSAMBLAJE .....	102
FIGURA 37. SITIOS DE CORTE ECORI Y PSTI PARA EL BIOBRICK U22490.....	102
FIGURA 38. SITIOS DE CORTE ECORI Y PSTI PARA EL PLÁSMIDO PSB1A3 .....	103
FIGURA 39. UNIÓN DEL BIOBRICK U22490 CON EL PLÁSMIDO PSB1A3 .....	104
FIGURA 40. ANOTACIONES FUNCIONALES Y ESTRUCTURALES .....	105
FIGURA 41. VISUALIZACIÓN DEL ENSAMBLAJE .....	106
FIGURA 42. RUTA METABÓLICA ENTRE LA LACTOSE Y EL PIRUVATO.....	111
FIGURA 43. VISUALIZACIÓN DEL CAMINO MOSTRADO EN LA TABLA 6.....	114
FIGURA 44. VISUALIZADOR DE ENSAMBLAJES .....	122
FIGURA 45. NÚMERO DE CAMINOS ANALIZADOS .....	125
FIGURA 46. NÚMERO DE CAMINOS ENCONTRADOS .....	126
FIGURA 47. VISUALIZACIÓN DE GRAFOS CON ALTA DENSIDAD .....	128
FIGURA 48. ERROR DE TRADUCCIÓN EN KEGGTRANSLATOR .....	129
FIGURA 49. CONEXIÓN ENTRE EL D-GLYCERALDEHYDE-3P Y PYRUVATE EN EL MAPA ECO01100 .....	132
FIGURA 50. BIOBRICKS NO DOCUMENTADOS .....	133

FIGURA 51. ELEMENTOS DE LAS RUTAS METABÓLICAS EXCLUIDAS DEL GRAFO ..... 135

## INDICE DE TABLAS

TABLA 1. COMPARACIÓN DE LOS REPOSITORIOS DE RUTAS METABÓLICAS.....	51
TABLA 2. RUTA METABÓLICA ENTRE LA LACTOSA Y EL PIRUVATO.....	82
TABLA 3. UNIPROT IDS PARA LAS PROTEÍNAS SELECCIONADAS .....	89
TABLA 4. REFERENCIAS CRUZADAS EN GENBANK PARA CADA PROTEÍNA.....	93
TABLA 5. TECNOLOGÍAS UTILIZADAS.....	108
TABLA 6. CAMINOS A PARTIR DE 101 RUTAS METOBÓLICAS .....	112
TABLA 7. CARACTERIZACIÓN DE GENES ENCONTRADOS.....	119
TABLA 8. CATÁLOGO COMPLETO .....	120
TABLA 9. CAMINOS ENCONTRADOS HASTA PROFUNDIDAD 16 .....	123
TABLA 10. HARDWARE PARA PRUEBAS DE RENDIMIENTO.....	127

# **CAPITULO 1. DEFINICIÓN DEL PROYECTO**

## **1.1. Planteamiento del Problema**

La creación de biocombustibles con bajo impacto ambiental en la actualidad es de vital importancia, debido a los altos índices de contaminación que se perciben en el país y el mundo. Por medio de técnicas desarrolladas en el campo de la Biología Sintética, es posible diseñar nuevos organismos que habilitan la producción de biocombustibles puros y de mejor calidad. Esta solución evita que estos sean mezclados con otros combustibles contaminantes, técnica que usualmente se utiliza para mantener su compatibilidad con las tecnologías existentes (Chin, 2008). Además, minimiza aún más los desechos nocivos que producen y que diariamente afectan a toda la población.

Otra ventaja de este proceso de reingeniería es que permite el aprovechamiento de materias orgánicas que tradicionalmente no se utilizan en la producción de biocombustibles. El suero de la leche, por ejemplo, representa una amenaza potencial de contaminación a las fuentes de agua cercanas a las plantas de producción de la industria láctea (Morales & Gurza, 2002), en caso de que este no sea utilizado.

Existen herramientas que dan soporte al proceso de diseño y ensamblaje de nuevos organismos, sin embargo, muchas veces, los profesionales en el área no cuentan con el conocimiento experto en el

área de estudio (Orozco Solano, 2012), que les permita tener una curva de aprendizaje sencilla.

Por otro lado, los sistemas actuales son de propósito general y no se especializan en un tema en específico como lo es la producción de biocombustibles, un ejemplo de ello, es el proceso de recolección de los componentes genéticos adecuados para alcanzar la degradación de materias orgánicas y la inferencia de funcionalidades de un organismo que ha sido rediseñado a partir de un conjunto determinado de piezas seleccionadas. Por lo anterior, se invierte una cantidad significativa de tiempo en la preparación de entradas para la creación de un nuevo diseño. La principal razón es que los datos se recolectan y se correlacionan manualmente basados en las fuentes de información disponibles y la experiencia desarrollada en el campo.

Además, en el proceso de diseño no se realizan análisis funcionales ni estructurales de los nuevos organismos diseñados a partir de los componentes utilizados y sus características, de modo que asistan al proceso experimental posterior a esta etapa. Las caracterizaciones estructurales permitirían describir el resultado final de manera concreta, por medio de la cuantificación de sus componentes genéticos. Por otro lado, las caracterizaciones funcionales permitirían realizar una aproximación sobre el comportamiento potencial del



diseño final, con el fin de realizar un mejor acercamiento desde un punto de vista funcional.

A partir de los puntos mencionados anteriormente, es posible caracterizar y resumir el problema de la siguiente manera, describiendo el mismo de lo más general a lo más específico:

- Existe la necesidad de la creación de biocombustibles de calidad y bajo impacto ambiental, lo que minimizaría una fuente más de contaminación y permitiría el aprovechamiento de desechos orgánicos no tradicionales.
- A través de la Biología Sintética es posible llevar a cabo el rediseño de organismos existentes, lo que permite la transformación de desechos orgánicos en biocombustibles.
- Estos diseños son asistidos por herramientas de propósito general, las cuales, no proporcionan información adicional a los usuarios con respecto al producto final.

A través del diseño de una herramienta para la asistencia del proceso de diseño es posible no sólo soportar al profesional que trabaja en el área, sino que también realizar un aporte al sector involucrado en la producción de biocombustibles en el país.

## **1.2. Objetivos**

### **1.2.1. Objetivo General**

Diseñar una herramienta prototipo que asista el proceso de diseño sintético de la bacteria E. Coli en la producción de biocombustibles por medio de la búsqueda y clasificación de componentes genéticos afines en catálogos y la caracterización de los nuevos ensamblajes.

### **1.2.2. Objetivos Específicos**

- Especificar el catálogo de componentes genéticos requeridos en el ensamblaje sintético de la bacteria E. Coli en producción de biocombustibles a partir del análisis y transformación de sus rutas metabólicas en estructuras computacionales y la aplicación de algoritmos de búsqueda.
- Establecer las características estructurales de los nuevos organismos ensamblados basados en su composición, cuantificando los componentes de los ensamblajes finales.
- Establecer una aproximación de las funciones de los nuevos ensamblajes sintéticos en la bacteria E. Coli a partir de las características de sus componentes, las cuales serán extraídas de repositorios de datos externos.

- Mostrar los ensamblajes sintéticos de la bacteria E. Coli por medio de computación gráfica.

### **1.3. Justificación**

El presente proyecto se desarrollará con el fin de proveer una herramienta que asista a los científicos experimentales en el diseño sintético de nuevos organismos, ya que las herramientas de diseño actuales para este fin son de propósito general, tales como la sección 2.4, y tienden a ser complejas para quienes no se encuentran familiarizados con los temas presentes (Orozco Solano, 2012). Además, se enfocan solamente en la construcción y no en la finalidad del producto que se quiere obtener.

La importancia de un sistema de este tipo reside en la posibilidad de realizar reingeniería de organismos involucrados en la producción de biocombustibles. Esto es posible a través del mejoramiento de su estructura, para lograr un mayor aprovechamiento de los desechos orgánicos que se generan en nuestro país. Por ejemplo, en la empresa láctea se desechan anualmente más de 675 mil toneladas del suero que proviene de la leche, ya que por su composición química son difíciles de tratar y en consecuencia, dicho suero es desechado en los ríos (Alvarado, García, & Rodríguez, 2012).

No obstante lo anterior, utilizando la Biología Sintética, es posible reutilizar el suero de la leche transformándolo en biocombustible. Esto alimentaría tecnologías a nivel nacional que utilizan como insumo combustibles fósiles, una de las mayores fuentes de contaminación ambiental.

Por medio de un sistema sencillo y orientado a un fin específico, se puede simplificar el proceso de diseño y ensamblaje, asistiendo a los profesionales del área en el diseño de nuevos organismos de una manera más simple y rápida. De esta manera, se contará con mayor información sobre el producto meta a construir. El alcance de la herramienta se enfocará desde la creación de un catálogo a partir del análisis y transformación de las rutas metabólicas de los distintos componentes genéticos hasta la caracterización de las secuencias ensambladas a partir de sus elementos. Esto contribuirá a que el proceso de diseño de organismos sintéticos sea más preciso en aplicaciones tales como la producción de biocombustibles, no sólo en el país sino en el nivel internacional.

Además de lo señalado, se apoyará la creación de nuevos biocombustibles no existentes en el país y la reutilización de materias orgánicas no tradicionales, que beneficiarán a la comunidad científica involucrada en el tema, a la población en general y al medio ambiente, reduciendo los índices actuales de contaminación.

El proyecto se llevará a cabo bajo la guía del profesor e investigador Allan Orozco, de la Universidad de Costa Rica (Escuela de Medicina y Postgrado de Ciencias Biomédicas) y la colaboración de ingenieros en Biotecnología y Biología del Instituto Tecnológico de Costa Rica y la Universidad Nacional. Lo anterior, con el propósito de alcanzar un diseño adecuado del sistema según las necesidades y la experiencia de los científicos experimentales en el área de la Biología Sintética y los biocombustibles.

#### **1.4. Hipótesis**

Con base en los recientes estudios en el campo de la Biología Sintética acerca de la producción de biocombustibles, los sistemas actuales utilizados en el ensamblaje sintético de organismos y las necesidades de la comunidad científica, es posible diseñar una herramienta prototipo que asista a los profesionales en el área de la Biotecnología y afines, en el diseño de nuevas funcionalidades y en el ensamblaje sintético de la bacteria E. Coli para producción de biocombustibles por medio de la transformación de las rutas metabólicas en una estructura de datos computacional adecuada que permita los análisis requeridos y la validación por criterio experto del diseño a través del desarrollo de un prototipo.

## **CAPITULO 2. FUNDAMENTACIÓN TEÓRICA**

### **2.1. La Bioinformática en Costa Rica**

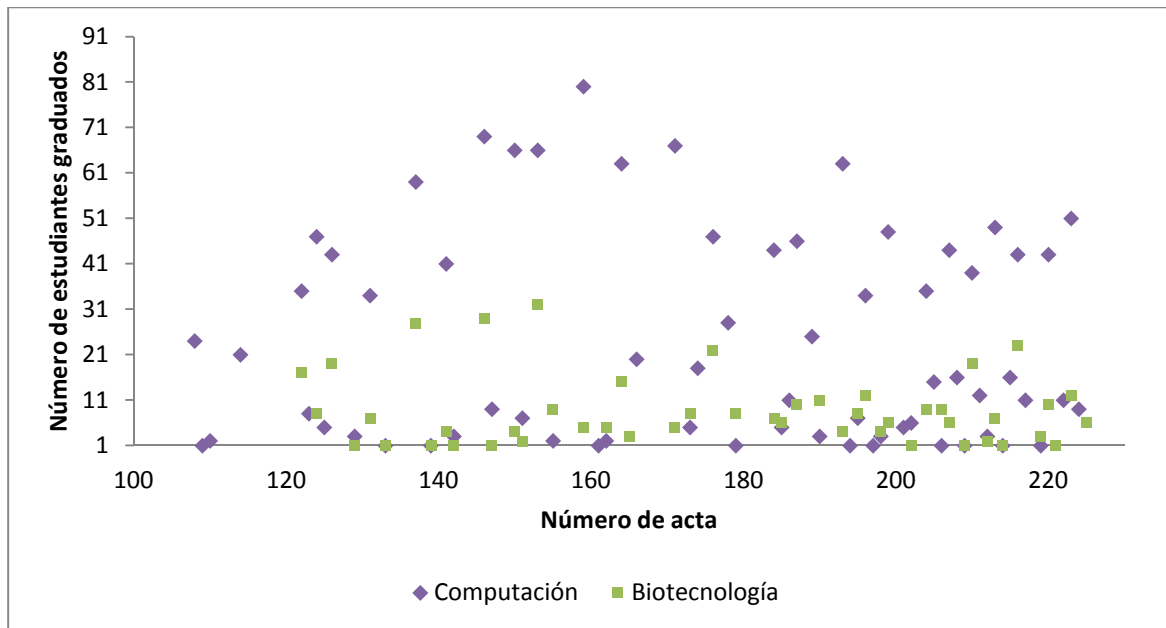
La Bioinformática es una ciencia interdisciplinaria que abarca principalmente conocimientos del área de la Biología y de Ciencias de la Computación. En esta disciplina, se procesan una cantidad significativa de datos e información y existe una necesidad importante de utilizar recursos computacionales para procesar y obtener resultados relevantes a partir de muestras amplias. Hace unos años, por ejemplo, secuenciar el genoma humano, el cual contiene  $3.2 \times 10^9$  pares de bases ( $\sim 3$  GB), podía tardar días y su costo se medía en millones de dólares. Ahora, con la ayuda de equipos especializados, se ha logrado optimizar esta tarea no sólo reduciendo su tiempo en horas (Kollewe, 2012), sino también reduciendo su costo significativamente en menos de \$1000 por un genoma completo (Perkel, 2013).

En Costa Rica, la Bioinformática ha surgido principalmente del ámbito académico e investigativo (Orozco, Morera, Jiménez, & Boza, 2013). Instituciones como la Universidad de Costa Rica (UCR), la Universidad Nacional (UNA), el Instituto Tecnológico de Costa Rica (ITCR), el Centro Nacional de Alto Tecnología (CeNAT), el Instituto Nacional de Biodiversidad (INBio), el Instituto Clodomiro Picado (ICP) y la Universidad Earth comenzaron a dar los primeros cursos y talleres de

Bioinformática y Biología Computacional desde el año 2000 (Orozco Solano, 2012).

Además, actualmente las universidades públicas en el país capacitan una cantidad significativa de profesionales al año en estas áreas afines, las cuales tienen un alto potencial para colaborar en el desarrollo del área de la bioinformática. Un ejemplo de ello, es el Instituto Tecnológico de Costa Rica (TEC), en donde se comparó y se evaluó la cantidad de graduados en las carreras de Ingeniería en Computación e Ingeniería en Biotecnología desde el año 2000 hasta el año 2013 a nivel de Bachillerato, a partir de la información brindada por el departamento de Admisión y Registro de dicha institución.

Dado a que la fundación de la carrera en Biotecnología se dio en 1997 (Aguilar, 2007), los primeros graduados se dieron a partir del año 2002. Desde ese momento hasta la fecha, el país cuenta con personal altamente calificado en ambas carreras, como se muestra en la Figura 1.



**Figura 1. Número de graduados en las carreras de Ingeniería en Computación y en Biotecnología del TEC desde año 2000 al 2013<sup>1</sup>**

En el año 2012, se inician los estudios de postgrado a nivel de maestría en Bioinformática y Biología de Sistemas en la UCR (fundada por el asesor del presente trabajo), como formación complementaria que les permita a los profesionales de éstas áreas desenvolverse mejor en el medio (Orozco Solano, 2012). También, en el 2004, el Consejo Nacional para Investigaciones Científicas y Tecnológicas (CONICIT), la Universidad de las Naciones Unidas (UNU) y el Programa de Biotecnología para América Latina y el Caribe (BIOLAC), firman un convenio con el fin de fortalecer el área de la biotecnología, por medio de talleres que permitieran la capacitación del recurso humano en esta

<sup>1</sup> Figura 1. Elaboración propia.



área para los países de Centroamérica y el Caribe (Consejo Nacional para Investigaciones Científicas y Tecnológicas, 2004).

La Bioinformática es aún una ciencia reciente e inmadura en el país, por lo que se enfrentan limitaciones que han frenado su crecimiento. Además, existen pocos profesionales capacitados en ésta área para satisfacer la demanda del mercado, ya que se necesitan conocimientos del área de la informática, ciencias biomédicas y biología (Orozco Solano, 2012) no sólo para dirigir proyectos, sino también para aprovechar las capacidades de las herramientas computacionales en los laboratorios experimentales.

Algunas instituciones, como el INBio, han desarrollado algunos sistemas de información como ATTA, un portal que permite consultar la base de datos de especímenes en el país en la Web, apoyando así la creación y divulgación de información sobre la biodiversidad (INBio, 2011). Sin embargo, la mayoría de enfoques en este Instituto se involucran solamente en el desarrollo de sistemas aplicados y no realmente en la creación de aplicaciones bioinformáticas.

Por otro lado, en el Centro Nacional de Alta Tecnología (CeNat), se han desarrollado proyectos como la predicción estructural de proteínas por medio de simulaciones computacionales. Estos proyectos están enfocados en realizar predicciones a partir de la estructura terciaria de la proteína, la secuencia de aminoácidos y utilizando el clúster de

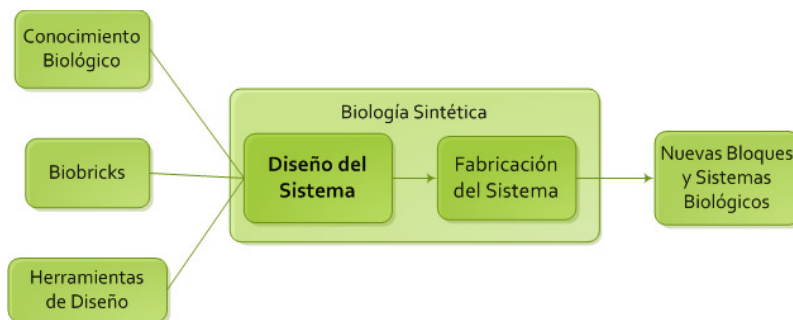
procesamiento del CNCA (Colaboratorio Nacional de Computación Avanzada). También hay proyectos más enfocados en almacenar gran cantidad de datos y no en su procesamiento, como el mantenimiento de genomas, que utilizan algoritmos de alineamiento de secuencias, clasificación, entre otros (Morales J. C., 2013).

Finalmente, la UCR dispone de un clúster de Bioinformática, en la Escuela de Medicina, donde se almacenan más de 350 servicios web alojados en la plataforma GALAXY. En este clúster se han programado las primeras aplicaciones para el análisis y alineamiento de secuencias a nivel local (Matarrita Araya, 2013), producto de los proyectos que se han llevado a cabo en la maestría de Bioinformática y Biología de Sistemas.

## **2.2. La Biología Sintética y los biocombustibles**

La Biología Sintética es el diseño y construcción de nuevos elementos y sistemas biológicos. Estos elementos son estandarizados en lo que se denominan biobricks. Esta ciencia además abarca el rediseño de sistemas biológicos naturales para propósitos útiles (Synthetic Biology Community, 2013). Mediante la ingeniería de nuevos componentes genéticos, que no se encuentran presentes en la naturaleza, se crean nuevas funcionalidades que reprograman el código genético de organismos ya existentes. Existen bacterias, por ejemplo, que

funcionan como pequeñas fábricas de biocombustibles y al ser modificadas genéticamente, mejoran su eficiencia. Esto evita mezclas con otros tipos de combustibles contaminantes y disminuyen el impacto que usualmente causan en los motores actuales (Chin, 2008). Cuando se crea una nueva pieza, se lleva a cabo un proceso de estandarización que permite abstraer su composición y a la vez definir interfaces de conexión que reducen su complejidad y permite que el conocimiento se comparta entre las diferentes organizaciones (Heinemann & Panke, 2006). Mediante estas ideas se simplifica el ensamblaje de un nuevo organismo por medio de bloques o "biobricks", proceso equivalente a armar las piezas de un rompecabezas. En la Figura 2, se muestran las entradas, salidas y procesos involucrados en la Biología Sintética.



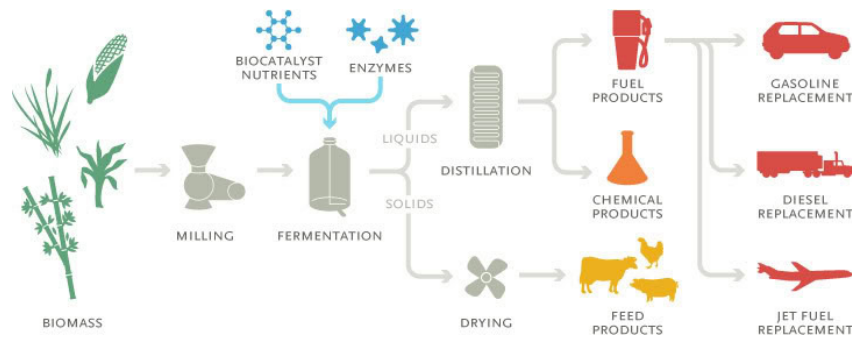
**Figura 2. Biología Sintética (Heinemann & Panke, 2006)**

La etapa más relevante para esta investigación es el diseño del sistema, ya que se considera que es factible dar soporte a los usuarios

en el ensamblaje de nuevos componentes genéticos, con el fin de obtener mayor información sobre el funcionamiento del producto final. Desde hace algunos años, la Biología Sintética se ha identificado como área de desarrollo y de oportunidades, al plantear la posibilidad de diseñar y aplicar ingeniería en organismos existentes. Las aplicaciones de esta metodología son diversas y una de ellas es la producción de biocombustibles.

Los biocombustibles son un tipo de combustible, líquido, gaseoso o sólido, derivados de materiales biológicos o biomasa (The Institute of Grocery Distribution, 2007). La principal diferencia con respecto a las fuentes tradicionales de energía, es que se derivan de recursos renovables.

En el mercado estadounidense, la mayoría de biocombustibles se clasifican en dos categorías: el etanol que se obtiene principalmente del maíz y el biodiesel que se produce a partir de los ácidos grasos. Sin embargo, ambos funcionan como aditivos en la gasolina o el diesel a menos que el motor se encuentre diseñado específicamente para utilizar este tipo de combustibles (Feltman, 2013).



**Figura 3. Proceso de la producción de bioetanol (Chin, 2008)**

Dependiendo del tipo de combustible y la materia orgánica que se utilice, el proceso puede variar, pero se mantienen los principios básicos.

En la Figura 3, se muestra el flujo para la producción de bioetanol. Inicialmente se selecciona y se muele la materia orgánica de origen vegetal, luego, se da la fermentación de azúcares y almidones contenidos en la biomasa por medio de microorganismos que producen etanol. Los subproductos son separados y los líquidos son destilados para eliminar la mayor cantidad de agua, de modo que pueda usarse como combustible. A pesar de las variaciones que pueda haber en el proceso, el principal interés se encuentra en el papel que juegan estos pequeños organismos y las reacciones que causan durante el procesamiento de los productos primarios.

Dependiendo de la materia orgánica que se utilice, se pueden crear distintos tipos de biocombustibles (Demirbas, 2008). A continuación se listan algunos de los biocombustibles más comunes:

- Bioetanol: derivado de cultivos de cereales, como el trigo, remolacha, maíz, granos de soya y caña de azúcar. Este es utilizado como sustituto del petróleo (gasolina).
- Biodiesel: derivado de cultivos de semillas oleaginosas, como la palma aceitera, girasol o granos de soya. Este es utilizado como sustituto del diesel.
- Biogas: derivado de desechos orgánicos, ya sea de animal o de fuente municipal, comercial o industrial. Este es utilizado como sustituto del gas natural.
- Otros biocombustibles: según la ingeniería de los organismos y del proceso, también se pueden producir otros combustibles para motores como el metanol, propanol y butanol.

Por medio de la Biología Sintética, se puede modificar los organismos que participan en el proceso de degradación ya sea para aumentar la producción o para sintetizar otros tipos de combustibles que disminuyan el efecto nocivo que producen en el ambiente. En particular, los estudios recientes se han enfocado en la reingeniería de la bacteria *Escherichia Coli* (*E. Coli*) (Hu, 2013), utilizando cepas inofensivas comúnmente utilizadas en la industria de la biotecnología. La ventaja de este organismo es que es fácil de cuidar, tiene un alto grado de reproducción y existe mucha información al respecto, dado a que ha sido estudiado ampliamente (Liu & Khosla, 2010). Además,

esta puede ser modificada para producir muchos tipos de biocombustibles como el etanol, el hidrógeno y el biodiesel (Fuga & Collier, 2013).

Los investigadores de la Universidad de California (UCLA), proponen la producción de biocombustibles a partir de la glucosa (Chin, 2008), modificando genéticamente la bacteria E.Coli para que esta sea un sintetizador eficiente de biocombustible. La ventaja es que por medio de este proceso se pueden producir alcoholes (como el isobutanol) que poseen una densidad similar a la gasolina, sin ser tan volátil ni corrosivo como el etanol, provocando menor daño a los motores. Mediante este estudio, no sólo se logró producir diferentes tipos de alcoholes no típicos, sino que también se utilizaron otros compuestos orgánicos como la levadura (Atsumi, Hanai, & Liao, 2008).

En la Universidad de Exeter, en el Reino Unido, otro grupo de investigadores modificaron cepas de la bacteria E. Coli para producir biocombustibles muy similares al diesel, por lo que no necesita mezclarse con otros productos del petróleo. Lo más relevante es que no se enfocan en modificar el metabolismo actual de la bacteria, sino en diseñar e implementar nuevas rutas metabólicas para la producción de combustibles renovables y relevantes para la industria (Howard, Middelhaufe, & Moore, 2013).

De esta manera, se muestra como la Biología Sintética es un mercado emergente y fértil, que promete importantes avances en el futuro. Actualmente, países como China, Holanda, Estados Unidos, Gran Bretaña y Suiza poseen una estrategia de desarrollo alineada con la Biología Sintética, por otro lado, el sector privado se encuentra invirtiendo fuertemente para alcanzar un mayor desarrollo de esta área científica (Piacente, 2011).

### **2.3. La Biología Sintética en Costa Rica**

La Biología Sintética en Costa Rica y en general Centroamérica ha sido fuertemente impulsada en los últimos años, no obstante, es aún un mercado joven debido a diferentes razones legales, socioeconómicas, éticas, entre otras. En el país, por ejemplo, no existe una legislación robusta que permita proteger la confidencialidad de la información genómica de los ciudadanos (Salas, 2013), la gestión en la metagenómica de organismos y el impacto ambiental asociado. Sin embargo, los proyectos poco a poco avanzan y contribuyen al desarrollo de esta ciencia en el país.

Hasta la fecha, en el país se ha identificado sólo un proyecto dedicado al campo de la Biología Sintética. Este proyecto cuenta con la participación de un grupo de estudiantes y profesores de la Universidad Nacional y el Instituto Tecnológico de Costa Rica. En el



año 2012, el equipo participó en iGEM (International Genetically Engineered Machine), una competencia a nivel mundial de biología sintética. El proyecto, denominado Cibus 3.0, consiste en la producción de biocombustibles utilizando desechos de la industria láctea, utilizando el suero de la leche. El equipo logró crear 10 bloques o biobricks sintéticamente, que fueron recombinados con la bacteria *E. Coli* para la producción de biodiesel. Finalmente, estos nuevos componentes sintéticos fueron agregados a los registros oficiales (Alvarado, García, & Rodríguez, 2012).

Existen instituciones que se encuentran trabajando en la producción de biocombustibles, como el Centro Nacional de Investigaciones Biotecnológicas (CENIBiot), los cuales utilizan biomasa vegetal, como las microalgas, para la producción de etanol y butanol (Rosales, 2013). Sin embargo, sus metodologías se basan en métodos tradicionales para la producción de biocombustibles. Además, no utilizan técnicas del área de la Biología Sintética, como por ejemplo, análisis y procesamiento de datos relativos a rutas metabólicas para el rediseño de nuevos organismos que contribuyan a la creación de nuevos biocombustibles.

## **2.4. Herramientas de diseño**

Con el fin de acelerar el proceso de ensamblaje de nuevos componentes, los usuarios o profesionales en el campo, utilizan herramientas que les permiten crear o armar su diseño "in silico" o de manera digital para luego comprobar su comportamiento en el laboratorio.

A pesar de que existen herramientas muy completas para el diseño de nuevos ensamblajes, se requiere un conocimiento a profundidad del área. Esto limita a muchos ingenieros experimentales aprovechar las aplicaciones en Bioinformática para este fin, una problemática visible no sólo en herramientas de ensamblaje sino también muchas otras herramientas del área de las ciencias biológicas.

Por lo anterior, es necesario mejorar el uso de los recursos bioinformáticos de manera que habilite a los investigadores a encontrar, interactuar, compartir, comparar y manipular información importante de manera más efectiva y eficiente (Bolchini, Finkelstein, Perrone, & Nagl, 2009).

A continuación se describen algunas de las herramientas de diseño existentes que no sólo permiten ensamblar los componentes genéticos sintéticos, sino también combinar estos bloques con el genoma del organismo seleccionado:

- Gene Designer 2.0 es una herramienta visual que permite el diseño de segmentos de ADN sintéticos, sin limitarse a los componentes que se encuentran en la naturaleza. Además, es factible editar y combinar los elementos genéticos mediante una interfaz gráfica entre otras características (Villalobos, Ness, Gustafsson, Minshull, & Govindarajan, 2006).
- SynBioSS es un conjunto de herramientas de Biología Sintética para la modelación y simulación de construcciones genéticas. Este utiliza un registro estándar de partes biológicas e interfaces gráficas y de línea de comandos para ejecutar los algoritmos de simulación (Hill, Tomshine, Weeding, Sotiropoulos, & Kaznessis, 2008).
- GeneComposer es un programa para realizar ingeniería sintética asistida de secuencias de ADN. Este sistema realiza una integración completa de datos estructurales y de secuencia que facilita las tareas en la etapa de diseño (Lorimer, Raymond, & Walchli, 2009).
- Pathway Hunter Tool es una herramienta que soporta el proceso de recolección de piezas genéticas para los diseños. Este sistema se enfoca en el análisis de conectividad entre los compuestos y las enzimas presentes en las rutas metabólicas de los nuevos organismos (Rahman, Advani, & Schunk, 2005).

Herramientas como las mencionadas en el punto anterior, poseen un nivel de dificultad significativo, principalmente para quienes no pertenecen al área. Una de las principales razones se debe a que su diseño se centra principalmente en su propósito y no en la usabilidad o la curva de aprendizaje que podrían presentar los nuevos usuarios. En general, muchos de los usuarios afirman que la comunidad bioinformática provee datos precisos y de gran valor a través de sus herramientas, sin embargo, frecuentemente encuentran que sus interfaces a estos recursos son engorrosas en cuanto a su uso y navegación (Pavelin, Cham, & Matos, 2014).

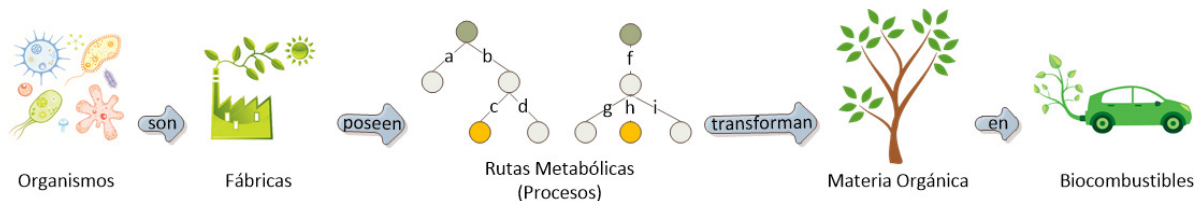
Dado estas afirmaciones, es importante que el diseño de la herramienta propuesta sea apropiada no sólo en cuanto a funcionalidad sino también en el despliegue de los resultados de manera simplificada y su presentación por medio de representaciones gráficas, que faciliten la comprensión de los usuarios.

Por otro lado, como se menciona en las secciones anteriores, no sólo existen recursos para el desarrollo de la bioinformática en el país, sino que también es posible innovar en área como la Biología Sintética por medio de proyectos interdisciplinarios que permitan la contribución y el enriquecimiento de los distintos proyectos activos y futuros en el área.

## CAPITULO 3. METODOLOGÍA

### 3.1. Descripción del proceso

En la naturaleza muchos organismos funcionan como pequeñas fábricas, es decir, a partir de la materia prima, se generan subproductos que son utilizados para otros fines en particular. Dentro de los organismos existen rutas metabólicas o procesos, que indican como los compuestos principales son degradados en compuestos más simples, que finalmente se descomponen en un producto meta. Un ejemplo de ellos, se muestra en la Figura 4, en donde los organismos que puede transformar la materia orgánica en subproductos, son la base principal para la producción de biocombustibles.



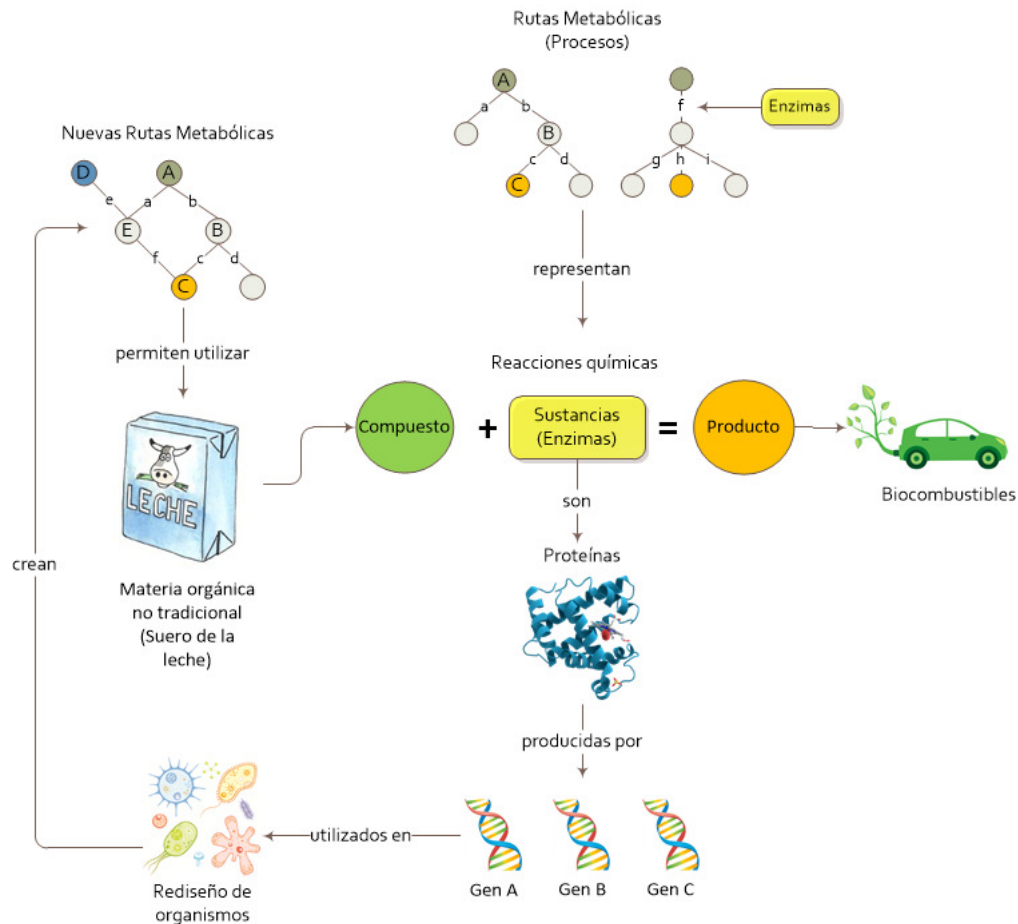
**Figura 4. Los organismos y sus procesos<sup>2</sup>**

Las materias tradicionales en la producción de biocombustibles suelen ser cereales o cultivos de semillas oleaginosas. Sin embargo, existen desechos orgánicos que no se reutilizan con el fin de producir biocombustibles. Una de las razones es que los organismos existentes

<sup>2</sup> Figura 4. Elaboración propia.

no poseen los procesos necesarios para realizar este tipo de transformación. Debido a esto, aquellos desechos que no tienen ninguna otra utilidad, se convierten en una fuente más de contaminación.

Las rutas metabólicas son una cadena de reacciones químicas catalizadas por sustancias que permiten llevar a cabo funciones de descomposición de materia orgánica, tal y como se muestra en la Figura 5.



**Figura 5. Diseño de nuevas rutas metabólicas<sup>3</sup>**

<sup>3</sup> Figura 5. Elaboración propia.

Los compuestos base son extraídos de la materia orgánica y los productos meta, obtenidos luego de una serie de reacciones, son utilizados como materia prima para producir biocombustibles. Estas sustancias catalizadoras son llamadas enzimas y pertenecen a uno de los grupos de las proteínas. Las proteínas pueden ser producidas por distintos genes y dependiendo del organismo, sus características pueden variar.

Las rutas metabólicas pueden ser mejoradas para alcanzar un propósito en específico, utilizando genes de otros organismos. De esta forma, se puede crear una selección de genes y luego modificar la ruta existente para lograr nuevas funciones. Las rutas de la bacteria E. Coli, por ejemplo, pueden modificarse para procesar desechos orgánicos como el suero de la leche. Mediante estas construcciones, se pueden producir los mismos substratos, como los ácidos grasos para la producción de biocombustibles, al igual que se creaban a partir de otras entradas, con la diferencia que los genes pueden pertenecer a otras rutas metabólicas.

### **3.2. Definición de etapas**

Con el fin de mejorar la eficiencia en la modificación de procesos tradicionales para la transformación de compuestos y productos, se propone la incorporación de elementos computacionales en la etapa de

diseño de rutas metabólicas. Esto permitirá no sólo la automatización de tareas sino que llevará a cabo análisis para la selección genes en el diseño de nuevas rutas metabólicas. A continuación se citan las etapas en las que se descompone la solución:

- Etapa 1: Análisis de rutas metabólicas.
- Etapa 2: Selección de proteínas.
- Etapa 3: Selección de genes.
- Etapa 4: Creación de un catálogo y sus ensamblajes.
- Etapa 5: Visualización de ensamblajes.

### **3.2.1. Etapa 1: Análisis de rutas metabólicas**

Para cada organismo, existe un gran banco de rutas metabólicas que describen los procesos internos de cada uno de ellos. En el campo de la Biología Sintética, es relevante conocer las posibles relaciones que se pueden establecer entre organismos, ya que a partir de ellas es posible crear nuevas propuestas para la creación de diseños que no existen en la naturaleza. Tradicionalmente, la mayoría de propuestas o herramientas existentes se enfocan en el análisis dentro de un sólo organismo, debido a que involucrar relaciones cruzadas entre distintos de ellos aumenta la complejidad en la construcción y en la validación. Con el fin de encontrar resultados relevantes con la información disponible en los repositorios de datos disponibles, se propone crear



una sola red de rutas metabólicas a partir de un subconjunto seleccionado. A partir de pequeñas redes, se localizan los puntos de enlace, representados en los elementos que tienen común y se conectan uno a uno hasta tener una sola red metabólica que permita realizar búsqueda de caminos de un punto a otro.

Antes de recorrer la estructura construida, se definen una lista de puntos de entradas y puntos de salida, o meta, la cual va ser dependiente de las necesidades de quién realiza los análisis. El mismo criterio aplicará para el subconjunto inicial de rutas metabólicas, lo que permitirá hacer una búsqueda más enfocada por tipo de ruta u organismo, eliminando ruido que puede agregar la utilización de rutas no relevantes para el producto final.

Para la selección de caminos, se utilizará el enfoque de todos los caminos posibles, basados en restricciones de inclusión, exclusión, y profundidad.

Por medio del tamaño del camino se permitirá conocer la ruta que requiera la menor cantidad de componentes genéticos, pero no necesariamente la más relevante para el análisis que se llevará a cabo. Por otro lado, listar todos los caminos posibles brinda flexibilidad a la hora de seleccionar la ruta más apropiada. Sin embargo, dado a que la red completa puede constar de miles a millones de elementos, se deben definir restricciones que acorten la búsqueda, no sólo por el

costo computacional sino también por la relevancia de los resultados finales. Estas restricciones definirán si un camino es más interesante que otro. A continuación se listan las restricciones habilitadas:

- Profundidad o tamaño del camino.
- Exclusión de compuestos o proteínas.
- Inclusión de compuestos o proteínas.

Al definir todas las entradas necesarias, se realizará el cálculo de todas las posibles rutas disponibles. Dado a que el proceso de selección es dependiente a factores específicos del estudio que se lleva a cabo, el profesional en el área será el responsable de seleccionar cual será la ruta con la cual se va a trabajar en las siguientes etapas.

### **3.2.2. Etapa 2: Selección de proteínas**

En una ruta metabólica, se pueden encontrar elementos de distintos tipos, en general, se definen en compuestos y proteínas. Los compuestos se descomponen en compuestos más simples debido a las reacciones químicas que llevan a cabo las proteínas. Los compuestos son obtenidos del medio y las proteínas son parte de cada organismo, por esto, es que el siguiente paso consiste en la extracción de las proteínas presentes en la ruta. El objetivo final es encontrar los genes necesarios para construir la nueva ruta metabólica y para esto se requiere conocer las proteínas asociadas.

De la ruta seleccionada en la etapa anterior, se filtran todos los elementos que representan una proteína, y se excluyen aquellas que no cumplan con alguno de estos criterios:

- Familia taxonómica relevante. Por ejemplo: Bacteria.
- Entrada verificada, es decir, fue manualmente validada y no generada de manera automática, computacionalmente.

### **3.2.3. Etapa 3: Selección de genes**

Una vez obtenida la lista de proteínas, se puede obtener la lista de genes responsables en su fabricación a partir de la información disponible en la estructura interna de las rutas metabólicas. Usualmente, estas contienen identificadores que permiten obtener más detalles sobre sus elementos y sus dependencias.

Es importante recalcar que la producción de una proteína no es única, sino que existen genes, de distintos organismos, que pueden producir la misma proteína, pero con algunas variaciones. Además, la selección de los genes es un proceso dependiente del organismo y el producto meta. Por este motivo, de ser necesario, el usuario se involucrará en la selección del gen o genes de interés, por medio de criterios basados en la literatura disponible que describe la función que ejecuta, los sitios de restricción presentes o características de los elementos que conforman la proteína (Alvarado Meza, 2014). De lo contrario, todos

los elementos encontrados serán agregados al catálogo de componentes genéticos.

Para cada gen, existe una cantidad de información significativa respecto a su origen, referencias, entre otros. Sin embargo, para el diseño de nuevos organismos, lo que se requiere es la secuencia de ADN codificante o CDS (Coding DNA Sequence) por sus siglas en inglés, la cual es la región del gen que codifica la proteína, como la que se muestra en la Figura 6.

### **E. coli chromosomal region from 76.0 to 81.5 minutes**

GenBank: U00039.1

[GenBank](#) [Graphics](#)

```
>gb|U00039.1|ECOYW76:1-1085 E. coli chromosomal region from 76.0 to 81.5 minutes
GATTTGGCATAGGCTTTTATTGGTCAGCGCCTGCCACGGTACTTTGACAAAATCCGGGTCGCCAAGATATT
CCGAGCGGTCCGGCTAGGCGTATTTCTCCGCTTCTGCCATGATTTGCATCGCATCGGCGCTGCCAAAGCC
GTATTTCTTCATATCGAAGTTTTCCAGAAATATTGAGGATTTGTACGATATGGATCCCGCCGGAGGATGGC
GGTGGCATGGAGTAAACCTGATACCCGCGATAATCGCCGCTTATCGGAGTGCCTTCGACCGCTTTATAGG
CTGCTAAATCTTCTTTAGTGATCAAGCCACCGTTTTTCTGCATCTCCTGGGCGATCTGTTCCGCAATCGT
GCCTTTATAGAATTCGTCCGGGCCGTTTTTCAGCAATCATCTCCAGGCTCTTTGCCAGGTTCCGCTGCACC
AGGGTGTCCGCTTTTTTCAGCGGCTCGCCCTCTTTCCAGAAGATAGCTTTACTGTTTTCGTGATTCGGCA
ACACTTCGCTACCGTAGGTTTTGAGATCGTCAGCCAGCGCGTCGTTAACGATAAAAACCATCGCGTGCCAG
TTTAAACGCGGGCTGCACGACTTTGTTTCAGCGGCATGGTGCCGTAATTTATCCAGCGCCAGCGAGAACT
GCTACCGTACCCGGTGTGCCGGAAGCCAGATGCCAAGTGAGTGATTTTTTGCTGTCCGGGTTGCCCTGAT
CATCGAGGAACATATCGCGGGTTCGCTTTGGCGGGTGCCATTTTCGCGAAAATCGATAGCCGTGGTATTGCC
ATTTTTCGAGCGGATTAACATAAAACCACCACCGCCAGATTCCCTGCCTGCGGATGCGTTACCGCCAGC
GCGTAGCCACCGCCACGGCGGCATCAACGGCATTCCCGCCCTCCTTGAGAATATCCACCCCCACCTGAG
TGGCAGTGGCGTCCACAGACGCTACCATTCCCTGTTTCGCGCGTACCGGGTGGAAAGACATCTTCTCCAC
ACCATACGAGACGGGCGGCGCAGGAGGCGGGCGGCGGCGCTAAAACAACCTTCTGAGAGCAGAGCAGCA
ATGGCCACCCGGCGTAAAAACGTTCGTTTTATCAT
```

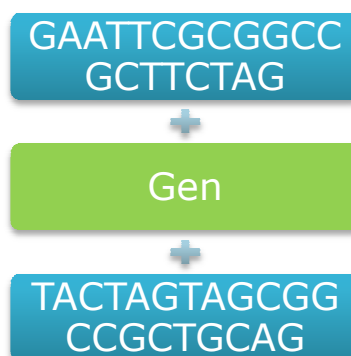
**Figura 6. Región codificante para el gen GMP1 (National Center for Biotechnology Information, 1994)**

El catálogo de biobricks, tendrá elementos que se construyen utilizando el CDS como base. Una vez conformadas estas piezas genéticas, será posible crear y visualizar los ensamblajes sintéticos.

Sin embargo, antes de que la pieza se encuentre lista para la siguiente etapa, es necesario verificar que se cumplan las siguientes restricciones:

- El gen tiene un CDS válido.
- El CDS no posee ningún sitio de corte o su complemento de las enzimas de restricción.
- El CDS es relevante para el usuario.
- La pieza posee los elementos estándar para constituir un biobrick. Para esto se deben agregar una serie de secuencias:

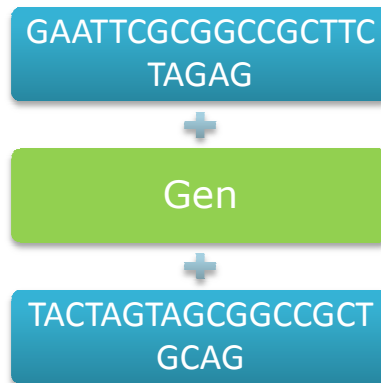
Si el gen es un CDS, se deben agregar el prefijo y sufijo de la Figura 7. En caso contrario, utilizará los prefijos y sufijos de la Figura 8.



**Figura 7. Prefijo y sufijo para una secuencia CDS<sup>4</sup>**

---

<sup>4</sup> Figura 7. Elaboración propia.



**Figura 8: Prefijo y sufijo para una secuencia no CDS<sup>5</sup>**

Estos prefijos y sufijos permiten que las piezas sean cortadas y ensambladas por medio de los procesos estándar establecidos por la fundación iGem. De esta forma, las piezas resultantes serán estandarizadas bajo el mismo formato que los biobricks.

#### **3.2.4. Etapa 4: La creación de un catálogo y sus ensamblajes**

Como se mencionaba anteriormente, existe un proceso estándar de ensamblaje de biobricks (International Genetically Engineered Machine Foundation, 2003) que permite el ensamblaje de biobricks utilizando técnicas tradicionales de clonación. Los ensamblajes definidos en iGem permite el ensamblaje de dos piezas ya sea dentro de otro organismo, como un plásmido o dos piezas entre sí. Un plásmido es una molécula de ADN que se replica y se transcribe de manera independiente al ADN

---

<sup>5</sup> Figura 8. Elaboración propia.

cromosómico. Tiene la ventaja que permiten la inserción de fragmentos de ADN y luego puede ser fácilmente introducido en una bacteria (Nature Education, 2013).

En el caso de la unión de dos piezas entre sí, éstas deben ser cortadas con enzimas que trabajan en determinados sitios de restricción (International Genetically Engineered Machine Foundation, 2003). Esto permite que los extremos sean compatibles y que tenga sentido la unión de sitios particulares.

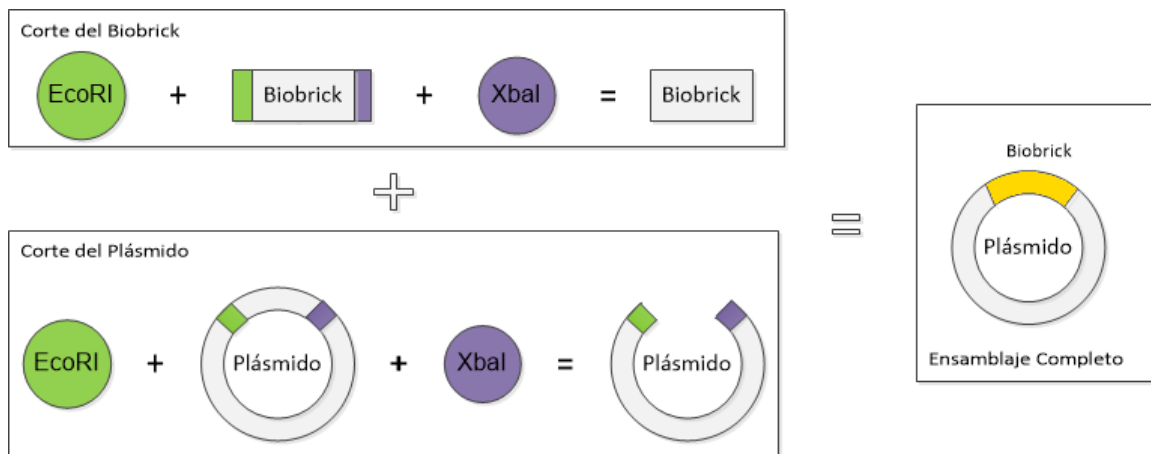
Como se menciona en la sección anterior, cada segmento va tener un prefijo y sufijo con sitios de restricción válidos. Esto va permitir cortar las distintas piezas a ensamblar por medio de las enzimas recomendadas en el estándar: EcoRI, XbaI, SpeI and PstI. Las enzimas EcoRI, XbaI definen sitios de corte en el lado izquierdo de la pieza y SpeI and PstI del lado derecho, como se muestra en la Figura 9.



**Figura 9. Biobrick estándar con su prefijo, sufijo y sitios de corte (International Genetically Engineered Machine Foundation, 2003)**

De esta forma, es posible realizar el ensamblaje en plásmidos, moléculas de forma circular, en los cuales se ensamblan estas nuevas piezas. Un elemento es compatible con otro, si para un sitio en

particular se utiliza la misma enzima. Es decir, para poder ensamblar un biobrick en un plásmido es necesario utilizar EcoRI en un extremo y XbaI en el otro como se muestra en la Figura 10. Al realizar un corte en el plásmido, es posible substituir el segmento sobrante por un biobrick y finalmente completar el ensamblaje.



**Figura 10. Ensamblaje estándar de biobricks<sup>6</sup>**

La metodología estándar puede aplicarse a varios casos, en particular se tomarán en cuenta los siguientes:

- *Pieza compuesta (2 Biobrick) + Biobrick:* se utiliza dos biobricks previamente unidos y se ensamblan a un nuevo biobrick.
- *Pieza compuesta (2 Biobrick) + Plásmido:* se utiliza dos biobricks previamente unidos y luego se ensamblan en un plásmido.

<sup>6</sup> Figura 10. Elaboración propia.



- *Biobrick + Plásmido*: se ensambla un biobrick individual dentro de un plásmido, como se muestra en la Figura 10.

Por otro lado, la construcción del catálogo se apoya en dos fuentes: los genes encontrados a partir de las búsquedas y análisis en rutas metabólicas y las categorías de interés del registro de biobricks (International Genetically Engineered Machine Foundation, 2003). Sin embargo, todos son tratados de la misma manera, ya que sin importar su procedencia, se definen los prefijos y sufijos necesarios para ser utilizado como un biobrick en caso de que no lo sea. Además, se contarán con un grupo de plásmidos previamente definidos.

Los ensamblajes se harán mediante una herramienta prototipo utilizando el catálogo propuesto, eliminando aquellos componentes que no cumplen con las restricciones definidas. Luego, se seleccionarán las piezas indicadas y las enzimas a utilizar dependiendo de la modalidad o caso que se escoja.

El resultado final será caracterizado estructuralmente y funcionalmente, con información correspondiente de las piezas seleccionadas.

### **3.2.5. Etapa 5: Visualización de ensamblajes**

Una vez completados los ensamblajes, estos se podrán visualizar mostrando sus detalles de forma gráfica y complementando con la información disponible de la construcción.

El producto final será visualizado en una representación circular y caracterizado estructural y funcionalmente a partir de los componentes utilizados.

### **3.3. Delimitación de entradas**

Con el fin de acotar el alcance del proyecto, se definirá un grupo limitado de entradas para llevar a cabo las tareas definidas.

Para la búsqueda en rutas metabólicas, se seleccionaron un conjunto de compuestos de entrada y salida, basados en la meta final que es la producción de biocombustibles. Las entradas se basan en compuestos comúnmente encontrados en desechos orgánicos idóneos para este fin, como por ejemplo, el suero de la leche o la broza del café. Por otro lado, los nodos terminales, son parte del proceso que produce energía en muchos organismos. Estos procesos a su vez, tienen un papel importante en la producción de biocombustibles. A continuación se listan los compuestos de inicio y final en la búsqueda de rutas:

#### Compuestos de entrada

- Lactosa
- Celulosa
- Quitina

- Colágeno
- Acetil Coenzima A
- Lignina

### Compuestos de salida

- Piruvato

Con respecto a las rutas metabólicas, existen distintos repositorios y herramientas que permiten la extracción de las mismas. Sin embargo, los formatos de salida y las herramientas para extraer la información son diferentes en cada caso. Con el fin de cumplir el objetivo de diseñar y validar la herramienta propuesta, se seleccionó uno de los repositorios.

Se analizaron las interfaces de datos de la Enciclopedia de Genes y Genomas de Kyoto, (KEGG) (Kanehisa Laboratories, 2014), Metacyc (Caspi, Altman, & Billington, 2014) y Reactome (Croft, O’Kelly, & Wu, 2011).

**Tabla 1. Comparación de los repositorios de rutas metabólicas<sup>7</sup>**

<b>Repositorio de Rutas</b>	<b>Acceso a Datos</b>	<b>Formatos de salida</b>
<b>KEGG</b>	Servicios Web REST FTP (Se necesita licencia)	KGML, Texto plano
<b>Metacyc</b>	Servicios Web REST APIs en Lisp, Perl, Java Búsquedas en mySQL locales	SBML, JSON, BioPax, FASTA, Texto plano
<b>Reactome</b>	Servicios Web REST (beta) Servicios Web SOAP Búsquedas en mySQL locales	SBML, SBGN, BioPax, PDF, Word, Protege

<sup>7</sup> Tabla 1. Elaboración propia.

Para la selección de la fuente de datos se tomó en cuenta no sólo los protocolos para el acceso a datos y los formatos de salida, sino la disponibilidad de librerías para la transformación de rutas metabólicas en estructuras de datos estándar como grafos. Además, se requería un acceso eficiente a las referencias de genes en otras bases de datos, específicamente las que se propusieron como parte de la implementación del prototipo. En la Tabla 1 se muestran los formatos de salida así como los accesos a datos de las principales bases de datos de rutas metabólicas.

Metacyc y Reactome fueron candidatos a considerar ya que ofrecen diferentes formatos de salida y un número significativo de protocolos para el acceso a datos. Sin embargo, KEGG era uno de los más utilizados por los usuarios y existían librerías confiables como KEGG Translator (Wrzodek, Dräger, & Zell, 2011) que permitían la traducción inmediata de las rutas metabólicas a formatos estándar en el ámbito computacional. Otro punto importante, es que posee referencias directas a bases de datos externas como UNIPROT, su interfaz de programación es simple y es de gran relevancia en el campo de la genómica para la comunidad científica.

Por estos motivos, las rutas metabólicas a recorrer se extraerán exclusivamente de KEGG, utilizando el organismo *Escherichia Coli* K-12 MG1655.

Con el fin de conectar y encontrar rutas relevantes entre los distintos compuestos en las rutas metabólicas, se definieron los siguientes algoritmos de búsqueda:

- Algoritmo de Dijkstra: este algoritmo permitirá encontrar una ruta entre dos compuestos de manera rápida y sencilla. A través de este mecanismo, será posible establecer las conexiones entre dos puntos como una implementación base, lo que definirá de primera entrada si se puede conectar dichos nodos antes de realizar una búsqueda de mayor complejidad.
- Algoritmo de búsqueda por anchura: dado a que para el usuario será relevante no sólo encontrar el camino más corto sino también analizar todas las posibles conexiones entre dos compuestos, es necesario realizar una búsqueda exhaustiva del árbol de rutas. Para esto se considerarán restricciones como exclusiones e inclusiones, que permitirán acotar la búsqueda y encontrar y filtrar los resultados de mayor relevancia para el usuario.

Para el ensamblaje de biobricks, se utilizarán los plásmidos y las enzimas que iGem define como estándar.

Los siguientes son los plásmidos seleccionados:

- pB1A3
- pSB1T3
- pSB3K3

Además, los sitios de restricción se limitan a:

- EcoRI
- PstI
- XbaI
- SpeI

Cada uno de estos sitios representan un patrón en particular, los cortes se realizarán en aquellos lugares donde se encuentre la secuencia o su complemento, utilizando los siguientes patrones y su reverso:

- EcoRI: GAATTC (CTTAAG).
- XbaI: TCTAGA (AGATCT).
- PstI: CTGCAG (GACGTC).
- SpeI: ACTAGT (TGATCA).

Los patrones cortarán la secuencia en un lugar en particular, lo que determinará cómo separar las cadenas de ADN. Este proceso se detalla en (International Genetically Engineered Machine Foundation, 2003).

En el catálogo también se incluye una lista de biobricks, como parte de las piezas candidatas a ensamblar. Se seleccionó la lista relacionada con la producción de combustibles, específicamente las siguientes categorías, tomadas del catálogo de biobricks.

- Registro de Partes Biológicas Estandarizadas (iGem).
  - Secuencias Codificantes.
    - Secuencias Codificantes de Proteínas.
      - Biosíntesis.
        - » Biosíntesis y degradación de AHL.
        - » Biosíntesis de los isoprenoides.
        - » Biosíntesis de los odorantes.
        - » Biosíntesis de los plásticos.
        - » Biosíntesis del butanol.
        - » Degradación del bisfenol A.
        - » Degradación de la celulosa.
        - » Otra enzimas misceláneas biosintéticas y de degradación.

Esta delimitación de entradas permite llevar a cabo el proyecto en el tiempo estimado, sin embargo, es importante mencionar las posibles extensiones y las implicaciones de cada una de ellas:

- Cambiar los nodos de entrada y salida: al agregar nuevos compuestos, se debe asegurar de seleccionar la muestra adecuada de rutas metabólicas de modo que existan caminos posibles entre los elementos seleccionados.
- Agregar nuevos repositorios de datos: Al incluir una nueva fuente de datos para las rutas metabólicas, se debe tomar en cuenta las conexiones a las bases de datos de proteínas y de genes. Esto permite mantener la relación entre rutas metabólicas, proteína y genes para el catálogo de biobricks.
- Utilizar varios repositorios de datos: dado a que cada repositorio tiene sus propios formatos de salida, es necesario crear un traductor de rutas metabólicas a un sólo formato estándar que permita manipularlas como estructuras de grafos.
- Agregar nuevos biobricks: debido a que la información de los biobricks no se encuentra estructurada, no existen filtros ligados al propósito final de los mismos. Es decir, agregar nuevas categorías no tiene ningún impacto en el sistema, basta con ingresar al sistema cuales biobricks se quiere agregar.

El objetivo principal del capítulo recién descrito será proveer al usuario una base y una guía sobre el proceso que se seguirá a alto nivel. En las siguientes secciones, se realizará una descripción más técnica, ligada a los detalles de implementación y específicamente a los



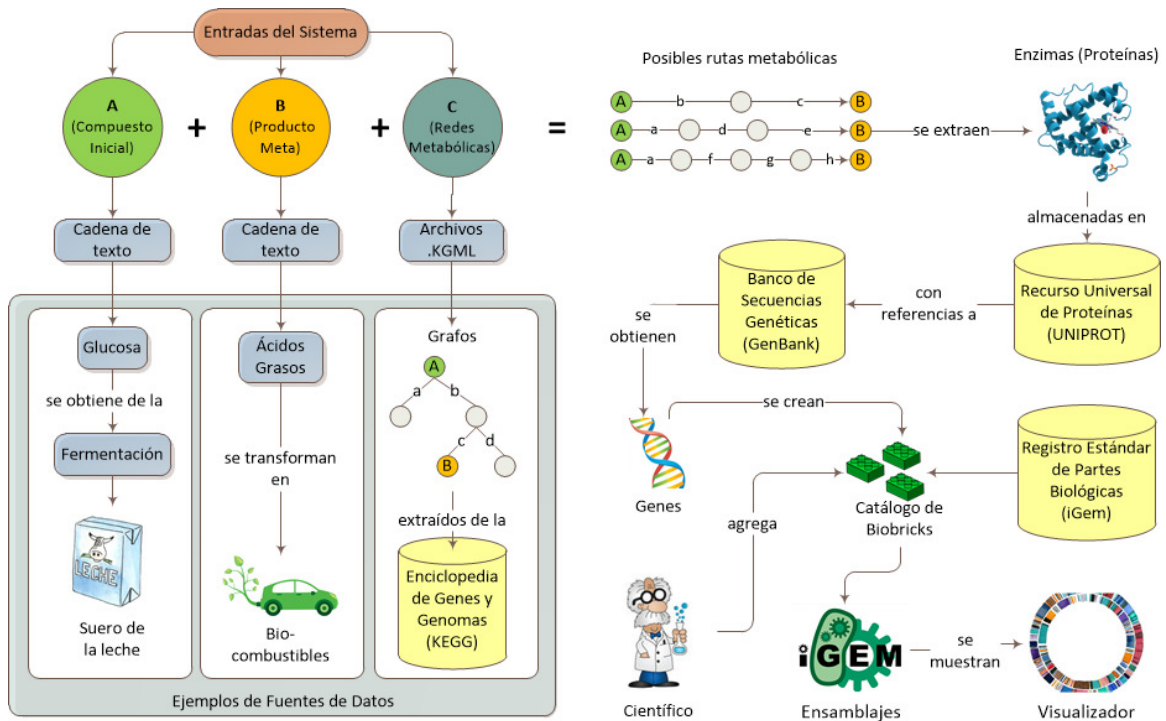
algoritmos, librerías y aplicaciones utilizados. Por otro lado, a través del desarrollo del prototipo, se validará el diseño propuesto con el fin de alcanzar las metas planteadas en la presente investigación.

## **CAPITULO 4. IMPLEMENTACIÓN DEL PROTOTIPO**

### **4.1. Diseño del Modelo Computacional**

En el proceso de rediseño de las rutas metabólicas, se necesita realizar un análisis profundo sobre las rutas existentes en las bases de datos. La descomposición de un compuesto en otro se puede hacer de diferentes formas y unas pueden tener mayor valor que otras. Además, algunos caminos pueden ser más largos o cortos y también pueden utilizar una cantidad variable de enzimas. Antes de poder seleccionar cuales son los genes se van a necesitar en la reingeniería de un organismo, primero se debe elegir la ruta que mejor se adapte a las necesidades del análisis y así extraer de la misma estos componentes genéticos necesarios para el ensamblaje.

En la Figura 11 se muestra el modelo computacional que define el alcance de la herramienta.



**Figura 11. Modelo Computacional<sup>8</sup>**

A continuación se detallan las diferentes etapas del modelo:

- a) Identificación de entradas: inicialmente, se necesita el nodo inicial y el nodo final con el fin de encontrar los posibles caminos entre ambos. Estas búsquedas estarán limitadas a una profundidad limitada.
  - Compuesto inicial: con el fin de delimitar el alcance los resultados, esta entrada deber pertenecer al grupo de los carbohidratos (azúcares).

<sup>8</sup> Figura 11. Elaboración propia.

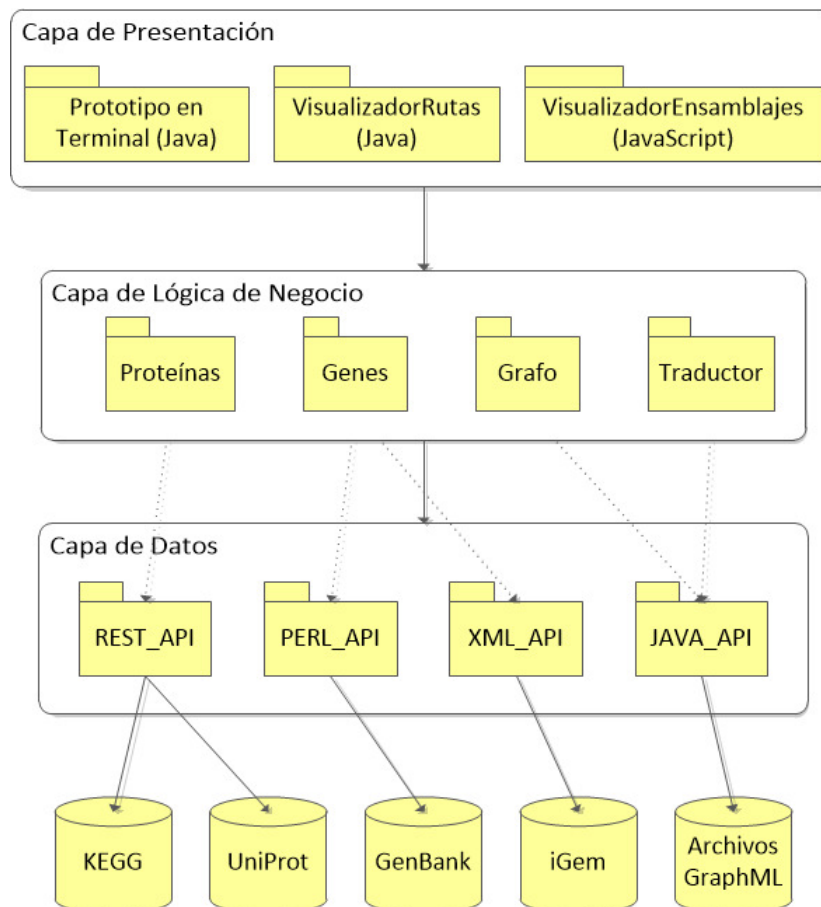
- Producto final: el producto final pertenecerá al grupo de ácidos grasos, como los triglicéridos. Estos son utilizados para la producción de biodiesel.
  - Redes Metabólicas: conjunto de grafos que describen los procesos de distintos organismos. Estos procesos o rutas metabólicas deberán contener el compuesto inicial y final en al menos una de ellas. Se utilizará como repositorio de rutas la base de datos KEGG, ya que posee estos mapas en formato XML y una interfaz de programación para que las aplicaciones puedan realizar consultas directamente.
- b) Cálculo de rutas metabólicas: una vez que se tienen los nodos de inicio y fin, se llevarán a cabo búsquedas en KEGG para calcular las posibles rutas que hay sobre los mapas disponibles.
- c) Selección de rutas: cada ruta se valorará por su longitud y en la cantidad de enzimas involucradas. Además, existirá la posibilidad de que el científico seleccione la mejor ruta de acuerdo a su criterio.
- d) Selección de genes: una vez que se selecciona la mejor ruta, se extraerán las enzimas involucradas en el proceso. Luego, se buscará información acerca de ellas en la base de datos "Universal Protein Resource (Uniprot)" en donde se extraerá cuales genes pueden producir estas enzimas. Una vez más, el

usuario elegirá cuál de ellos se apega mejor al análisis que se está llevando a cabo.

- e) Creación de biobricks: utilizando estos genes se realizarán búsquedas en iGem Registry para buscar coincidencias de los biobricks existentes. En caso de no existir, el usuario se encargará de construir el biobrick a partir de la información encontrada, ya que esto conlleva una serie de pasos adicionales fuera del alcance de este proyecto.
- f) Creación del catálogo en el visualizador: luego de realizar los análisis mencionados, los biobricks encontrados o construidos serán cargados en la herramienta de visualización como entradas.
- g) Visualización del ensamblaje: una vez generado el catálogo, la herramienta habilitará el ensamblaje de biobricks sobre la bacteria E. Coli. Los ensamblajes finales se caracterizarán de manera estructural y funcional, con respecto a las proteínas involucradas en el proceso.

Concretamente, el desarrollo del modelo computacional se realizó en un prototipo operado desde línea de comandos y las visualizaciones por medio de la generación de archivos HTML que muestran el resultado final de los ensamblajes.

En la Figura 12, se diagrama la arquitectura de la aplicación. De manera inicial, se definen las interfaces a cada una de los repositorios de datos. Luego, en la capa de lógica se agrupan los métodos de manera funcional y finalmente, se tienen tres interfaces en la capa de presentación: aplicación por línea de comandos, el visualizador de rutas metabólicas por medio de librerías en Java y el visualizador web para los ensamblajes en JavaScript.



**Figura 12. Arquitectura de la aplicación<sup>9</sup>**

<sup>9</sup> Figura 12. Elaboración propia.

## **4.2. Extracción de datos**

Las rutas metabólicas son una serie de reacciones químicas que ocurren dentro de una célula. A lo largo del tiempo, estas redes han sido estudiadas y documentadas en distintas bases de datos.

Para el desarrollo de este proyecto se seleccionó como principal fuente de datos, el repositorio de rutas metabólicas en KEGG, el cual contiene información para la comprensión de las funciones y utilidades de alto nivel del sistema biológico, cómo la célula, el organismo y el ecosistema. Estos datos son recolectados a partir de la secuenciación de genomas y otras tecnologías experimentales de alto rendimiento (Kanehisa Laboratories, 2014).

Las redes metabólicas se encuentran representadas en mapas (Apéndice 1), los cuales se pueden descargar de manera individual por medio de la interfaz de programación de la aplicación (API) bajo el esquema REST (Fielding, 2000). Los métodos disponibles en el API, se encuentran detallados en (Kanehisa Laboratories, 2014).

Estos mapas se encuentran en formato KGML (Apéndice 6), un formato de intercambio entre los mapas de rutas almacenados KEGG que permite el dibujo automático de rutas metabólicas. Además, provee facilidades para llevar a cabo análisis computacionales y modelos de redes de genes y/o proteína y redes químicas. Cada componente en el mapa, puede conectarse según su naturaleza,

creando enlaces entre sus reacciones, enzimas o genes (Kanehisa Laboratories, 2014).

Dado que para este proyecto se definió trabajar con la bacteria E. Coli, se utilizaron mapas específicos a este organismo, los cuales proveen enlaces basados en sus genes. A diferencia de los mapas genéricos, estos incluyen sólo la información relevante al organismo indicado, por lo que se evita la inclusión innecesaria de datos en los análisis relativos a otros organismos. Específicamente, se utilizarán los mapas bajo el código de organismo "eco", el cual corresponde a la bacteria Escherichia Coli K-12 MG1655.

Para la extracción de datos, a partir de una lista previamente definida, ya sea por el sistema o el usuario, se descargan los archivos de entrada por medio de scripts (Ver Apéndice 7).

La lista de mapas de interés se puede obtener a partir de búsquedas en KEGG por medio de palabras claves.

Por ejemplo, al realizar la búsqueda de los mapas que coinciden a la palabra "Lactose", se obtiene la siguiente lista:

- map00052
- map01110
- map00040
- map04973
- map00053
- map02030
- map00512
- map02010
- map04978
- map01100
- map02060
- map00600
- map00520
- map00051
- map00601



Una vez que se descarga el conjunto de archivos KGML, correspondientes a la lista de entradas, se procede a la etapa de transformación.

Para la recolección de datos, se corrieron los análisis sobre los mapas que coincidieran en la búsqueda de las palabras "Lactose" y "Pyruvate". Los resultados se muestran con un prefijo "map" dado que la búsqueda no se limita a un organismo en particular. Por esta razón, al asignar el código del organismo seleccionado ("eco" para la sepa E. Coli de interés), los mapas no coincidían a archivos KGML válidos. Dado a que el número de casos no era significativo y los mapas no estaban disponibles para este organismo, se excluyeron del conjunto de archivos que se utilizaron en el análisis. La

Figura 13 muestra un ejemplo de la dirección base utilizada para la descarga de archivos.

<http://www.kegg.jp/kegg-bin/download?entry=eco00010&format=kgml>

**Figura 13. URL estándar para la descarga de mapas de KEGG<sup>10</sup>**

---

<sup>10</sup> Figura 13. Elaboración propia.

### **4.3. Rutas Metabólicas**

Una vez descargados los mapas en un formato más flexible es posible llevar a cabo el análisis a partir de los mapas utilizados como datos de entrada. A continuación se describe el proceso de transformación y análisis en rutas metabólicas.

#### **4.3.1. Transformación de entradas**

Las rutas metabólicas representadas en los mapas de KEGG tienen un formato basado en XML, por lo que debe ser traducido e interpretado para poder crear una estructura de datos apropiada para su navegación. Además, dado a que están manualmente anotados, alguna información requiere ser complementada con referencias de otras bases de datos alternas.

Para llevar a cabo esta tarea, se seleccionó la herramienta KEGGtranslator (Wrzodek, Dräger, & Zell, 2011), de la Universidad Tübingen en Alemania, que permite la conversión de archivos KGML en múltiples formatos de salida. A diferencia de otros traductores, esta busca los componentes faltantes en algunas reacciones fragmentadas, soportando una variedad de formatos de salida, como por ejemplo: SBML, BioPAX, SIF, SBGN, SBML-qual, GML, GraphML, JPG, GIF y LaTeX. Esta herramienta lleva a cabo las traducciones por medio de

una interfaz gráfica o por línea de comandos, con distintas opciones disponibles para el análisis de los archivos de entrada.

Como formato de salida se seleccionó GraphML (GraphML Project Group, 2007), un lenguaje que permite la descripción de las propiedades estructurales del grafo. Este permite implementar grafos dirigidos y no dirigidos, representaciones gráficas entre otros. Otra ventaja, es que existen librerías que soportan este formato de forma nativa, creando una estructura robusta a partir de archivos bajo este formato.

Una vez definidas estas dependencias, se llevó a cabo la traducción de de las rutas metabólicas en formato KGML a formato GraphML, utilizando la herramienta KEGGTranslator. Además, se excluyeron todos los nodos que no tuvieran conexión, ya que al estar aislados tienen poca relevancia para los análisis.

#### **4.3.2. Construcción de rutas a través de grafos**

Debido a que el formato GraphML es utilizado naturalmente para la creación de grafos, existen varias librerías disponibles para su construcción. Esto permite abstraer la implementación de la estructura de datos y enfocarse en los algoritmos y análisis.

A partir de esta etapa, se trabajó con la librería Blueprints, del proyecto Tinkerpop (TinkerPop team, 2009), el cual consiste en una

colección de aplicaciones en el dominio de los grafos. En particular, Blueprints es una colección de interfaces e implementaciones que opera sobre un modelo que consiste en un grafo de propiedades (Rodríguez, 2012).

Este tipo de grafo consiste en una estructura de nodos y arcos, los cuales tienen una colección de propiedades definidas por un mapa de llaves y valores. Esto permite el manejo de propiedades variables, sin limitar su capacidad ni el momento en el cual se pueden definir.

Para la construcción del grafo, se cargan los archivos GraphML por medio de la librería en Java, creando así un grafo de propiedades como estructura de datos principal para la búsqueda de rutas metabólicas. La transformación de datos se muestra en la Figura 14.



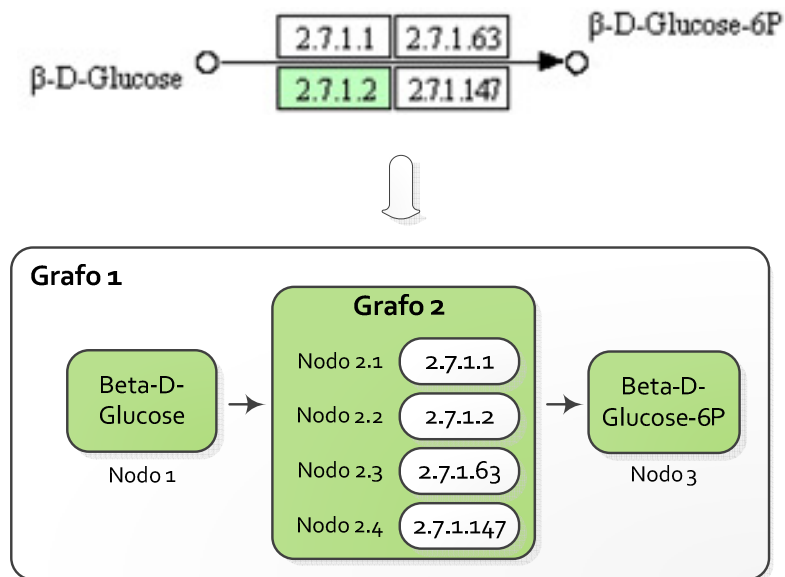
**Figura 14. Proceso de transformación de datos<sup>11</sup>**

Durante este proceso, se encontraron algunas limitaciones que debieron resolverse antes de continuar con la siguiente etapa. La principal limitación se debió a la aparición de grafos anidados, ya que es una característica que no es soportada por la librería de grafos Blueprints.

---

<sup>11</sup> Figura 14. Elaboración propia.

En las redes metabólicas es común encontrar varias enzimas presentes en el camino de un compuesto a otro. Al estar agrupadas en los mapas originales significa que cualquiera de ellas podría estar presente, representando un camino nuevo entre ambos puntos. Cuando esto es traducido al formato GraphML, el grupo de enzimas es representada en un mismo nodo y declarado como un grafo anidado, como se muestra en la Figura 15 con el Grafo 2.



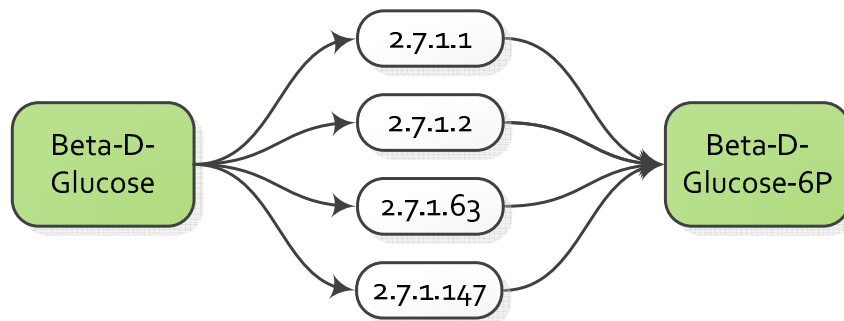
**Figura 15. Grafos anidados debido a la agrupación de enzimas<sup>12</sup>**

El Grafo 2 sólo es utilizado como un contenedor y no posee las mismas propiedades como el resto de los nodos. Esto causa que la librería falle

<sup>12</sup> Figura 15. Elaboración propia.

en la traducción, ya que no existe la propiedad correspondiente al identificador único del nodo.

Como solución, se desarrolló un script en Ruby que carga los archivos en una estructura de datos como un objeto XML. Luego, el algoritmo implementado se encarga de analizar cada nivel de la jerarquía y reubicar aquellos nodos que representan grafos anidados. Esto permite actualizar las referencias para construir caminos independientes para cada una de las enzimas como se muestra en la Figura 16.



**Figura 16. Grafos anidados transformados en caminos independientes<sup>13</sup>**

Una vez que se resuelven este tipo de inconsistencias en los datos de entrada, el proceso de traducción es aplicado a todos los archivos KGML que se encuentren disponibles al iniciar. Luego de completar este proceso, todos los grafos resultantes se unen en una sola estructura, la cual se describe bajo la siguiente definición:

---

<sup>13</sup> Figura 16. Elaboración propia.

Dados los siguientes conjuntos de grafos:

$$G_1 = \{b\}$$

$$G_2 = \{a, c\}$$

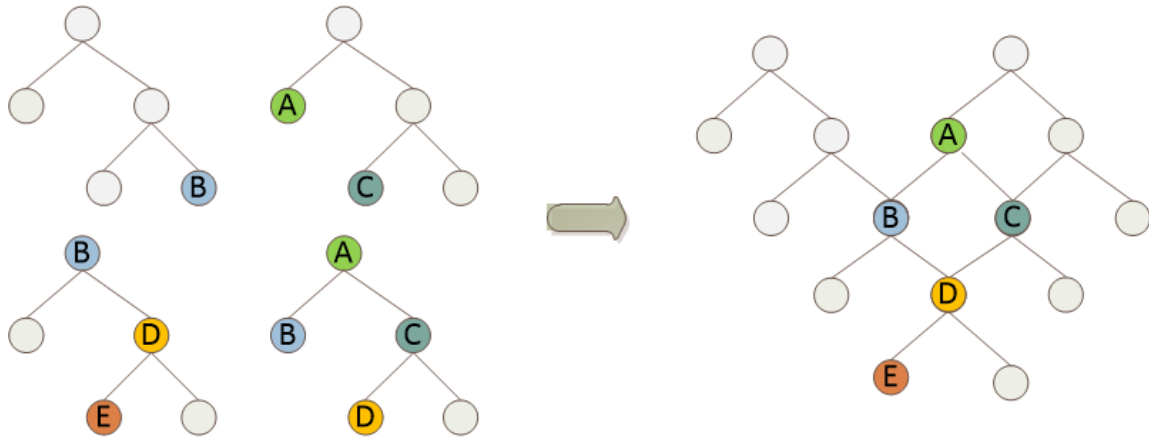
$$G_3 = \{b, d\}$$

$$G_4 = \{a, b, c, d\}$$

Es posible construir un sólo grafo, mediante la unión de estos subconjuntos:

$$H = \{G_1, G_2, G_3, G_4\}$$

Para la implementación de esta construcción, se toma cada nodo individual de los grafos existentes y se agregan a un nuevo grafo. Se valida que ninguno de ellos exista en la nueva estructura y de ser el caso, se establecen nuevas conexiones a los nodos existentes para cada nuevo elemento. Al completar esta etapa, el resultado será un grafo único que será utilizado para la búsqueda de caminos en la siguiente etapa. La Figura 17 muestra cómo al crear un sólo grafo, podrían habilitarse nuevas rutas como es el caso del camino de A → E.



**Figura 17. Unión de grafos independientes en una sola estructura<sup>14</sup>**

Una vez que los mapas se encuentran descargados, el usuario puede seleccionar un subconjunto de ellos para iniciar el procesamiento de rutas y la creación del grafo, como se muestra en la Figura 18.

<sup>14</sup> Figura 17. Elaboración propia.



```
=====
P A T H W A Y S   A N A L Y Z E R
Welcome to the Pathways Analyzer!
By Laura Vasquez
=====

Initializing.. Details will be logged here: PathwaysAnalyzer.log
[INFO] Loading Kegg info from file..
[INFO] Adding Plasmid: psb1a3
[INFO] Adding Plasmid: psb1t3
[INFO] Adding Plasmid: psb3k3
[INFO] Parsing biobricks file biobricks.list..
[INFO] Loading file part.BBa_C0060..
[INFO] Loading file part.BBa_C0060.1..
[INFO] Loading file part.BBa_C0061..
[INFO] Loading file part.BBa_C0070..
[INFO] Loading file part.BBa_C0076..
[INFO] Loading file part.BBa_C0078..
[INFO] Loading file part.BBa_C0083..
[INFO] Loading file part.BBa_C0160..
[INFO] Loading file part.BBa_C0161..
[INFO] Loading file part.BBa_C0170..
[INFO]
[INFO] [Step 1. Metabolic Pathways]

[USER] Please enter a list of pathways to be analyzed [pathways.list]. Use default? (y/n)
y█
```

**Figura 18. Carga de mapas, biobricks y plásmidos.<sup>15</sup>**

Al iniciar la aplicación, esta cargará los datos ya previamente definidos como la lista de plásmidos y biobricks. Luego el usuario deberá seleccionar la lista de mapas que se utilizarán en el análisis o también seleccionar el archivo definido por defecto.

Los mapas serán cargados o traducidos sólo si no han sido utilizados previamente, es decir, si el archivo se encuentra disponible en formato GraphML, se cargará inmediatamente como se muestra en la Figura 19. En caso contrario, se buscará el archivo KGML y se transformará

---

<sup>15</sup> Figura 18. Elaboración propia.

por medio del KEGGtranslator. En caso de que ninguno de los dos se encuentre disponible, se excluirá el mapa del análisis.

```
[INFO] Translating 102 pathways into graphs..  
[INFO] Parsing pathway eco00010..  
[INFO] No translation required for: data/graphml/eco00010.graphml  
[INFO] Loading file data/graphml/eco00010.graphml..  
[INFO] Graph loaded successfully => tinkergraph[vertices:90 edges:165]  
[INFO] Parsing pathway eco00020..  
[INFO] No translation required for: data/graphml/eco00020.graphml  
[INFO] Loading file data/graphml/eco00020.graphml..  
[INFO] Graph loaded successfully => tinkergraph[vertices:60 edges:101]  
[INFO] Parsing pathway eco00030..  
[INFO] No translation required for: data/graphml/eco00030.graphml  
[INFO] Loading file data/graphml/eco00030.graphml..  
[INFO] Graph loaded successfully => tinkergraph[vertices:81 edges:150]
```

**Figura 19. Carga de rutas metabólicas en formato GraphML<sup>16</sup>**

Una vez que todos los mapas disponibles se carguen satisfactoriamente, se unirán en el mapa final como se muestra en la Figura 20, creando un sólo grafo que permita hacer búsquedas entre los distintos componentes.

```
[INFO] Merging graph eco03060..  
[INFO] Merging graph eco03070..  
[INFO] Merging graph eco03410..  
[INFO] Merging graph eco03420..  
[INFO] Merging graph eco03430..  
[INFO] Merging graph eco03440..  
[INFO] Merging graph eco04122..  
[INFO] Pathways translation to graphs was completed.  
[INFO] Hypergraph: tinkergraph[vertices:4896 edges:6486]
```

**Figura 20. Unión de rutas metabólicas en un sólo grafo<sup>17</sup>**

---

<sup>16</sup> Figura 19. Elaboración propia.

<sup>17</sup> Figura 20. Elaboración propia.

Una de las limitantes al cargar la información desde los archivos KGML, fue que no todos los elementos tenían asociado un nombre, sino que sólo el identificador. Esto limitaba al usuario al no poder realizar las búsquedas a través de palabras claves, ya que algunos elementos se encontraban incompletos.

Como solución, se implementó un diccionario local que se completa con búsquedas directamente en la base de datos de KEGG. La primera vez que una reacción, por ejemplo, no posee un nombre significativo, se solicita la información por medio del API REST de KEGG como se muestra en la Figura 21.

```

ENTRY      R01678                      Reaction
NAME       Lactose galactohydrolase
DEFINITION Lactose + H2O <=> alpha-D-Glucose + D-Galactose
EQUATION   C00243 + C00001 <=> C00267 + C00124
REMARK     Same as: R06098
RPAIR      RP00437 C00124_C00243 main
           RP01684 C00243_C00267 main
           RP05702 C00001_C00124 leave
ENZYME     3.2.1.23             3.2.1.108
PATHWAY    rn00052 Galactose metabolism
           rn01100 Metabolic pathways
ORTHOLOGY  K01190 beta-galactosidase [EC:3.2.1.23]
           K01229 lactase-phlorizin hydrolase [EC:3.2.1.108 3.2.1.62]
           K12111 evolved beta-galactosidase subunit alpha [EC:3.2.1.23]
           K12112 evolved beta-galactosidase subunit beta
           K12309 beta-galactosidase [EC:3.2.1.23]

```

**Figura 21. Extracción de Datos de KEGG por medio del API REST<sup>18</sup>**

<sup>18</sup> Figura 21. Elaboración propia.

Con el fin de extraer la información de la solicitud a KEGG, se utilizó una variación del código utilizado en la base de datos UniProt (UniProt Consortium, 2012), cambiando la dirección del servidor a <http://rest.kegg.jp/get/>.

La implementación del diccionario se realizó por medio de un archivo de texto simple, de forma <llave, valor>:

- rn:R00127,ATP:AMP phosphotransferase
- rn:R03067,2-amino-4-hydroxy-6-hydroxymethyl-7
- rn:R05051,L-erythro-4-Hydroxyglutamate:NAD+ oxidoreductase
- rn:R03066,2-amino-4-hydroxy-6-hydroxymethyl-7

Al inicio del programa, los datos son cargados en la aplicación en una tabla hash, que permite la consulta eficiente de los componentes faltantes a la hora de autocompletar el grafo.

Luego de completar los grafos y su información, se procedió a llevar a cabo los algoritmos de búsqueda, los cuales se describen en la siguiente sección.

### **4.3.3. Algoritmos de búsqueda**

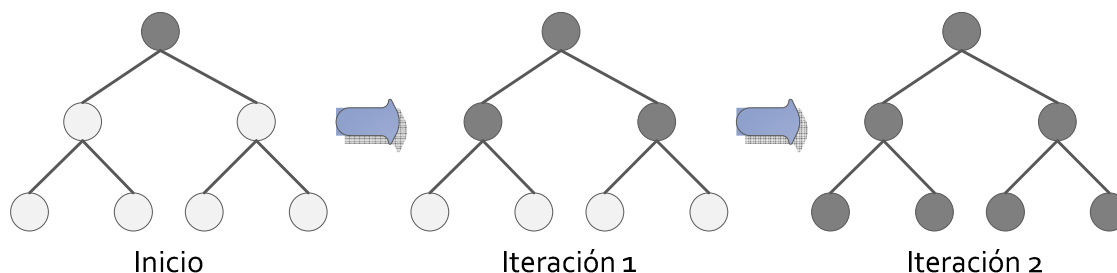
Al habilitar una estructura apropiada para crear relaciones entre los distintos nodos, es posible diseñar algoritmos de búsqueda que

permita navegar y construir nuevas conexiones a partir de una serie de reglas.

El objetivo principal consiste en encontrar todos los posibles caminos existentes entre dos nodos de entrada, dado un subconjunto de rutas metabólicas.

La librería utilizada para la implementación del grafo, provee algoritmos de búsqueda basados en JUNG (Java Universal Network/Graph Framework, por sus siglas en inglés), un marco de desarrollo que provee métodos para modelaje, análisis y visualización de los grafos (The JUNG Framework Development Team, 2010). Inicialmente, esta librería permitió llevar a cabo las validaciones iniciales utilizando las funciones disponibles para encontrar el camino más corto, como lo es el algoritmo de Dijkstra (Morris, 1998).

Una vez realizada la validación de la estructura, se realizó una implementación personalizada de un algoritmo que encontrará todas las rutas posibles, basados en la búsqueda por anchura (BFS, por sus siglas en inglés) y un sistema de restricciones que permite filtrar por prioridad los resultados encontrados. En la Figura 22 se muestra el algoritmos base de búsqueda por anchura.



**Figura 22. Búsqueda por anchura<sup>19</sup>**

La búsqueda por anchura consiste en la búsqueda del camino más corto, basada en la expansión de los sucesores de cada nodo y en la revisión exhaustiva de cada una de ellos hasta el siguiente nivel de profundidad. En cada iteración se revisará un nivel de sucesores hasta encontrar el nodo meta o hasta alcanzar el nivel de profundidad máxima definido.

Tradicionalmente, este algoritmo también es utilizado para encontrar el camino más corto. Sin embargo, dado a que se necesitan todos los caminos posibles, se implementa una pequeña variación utilizando el mismo esquema. Esta variación no agrega cada sucesor como un nodo pendiente a visitar, sino que se guarda el camino calculado hasta ese punto como una nueva opción y así para cada uno de los sucesores. Con esto, se tendrá una cola de caminos posibles, a los que iterativamente, se revisarán de nuevo los siguientes sucesores, creando así nuevas rutas en la lista en cada nivel.

<sup>19</sup> Figura 22. Elaboración propia.

Para encontrar todos los caminos posibles se requiere recorrer toda la estructura, lo que degrada rápidamente el rendimiento y el uso de la memoria con cada nivel que se analiza. Como una mejora que permita descartar aquellas rutas que no tienen ningún valor, se agregan las siguientes restricciones:

- La búsqueda será delimitada por una profundidad determinada.
- El usuario ingresará un conjunto de nodos a descartar, lo que reducirá el área de búsqueda del algoritmo.
- El usuario ingresará un conjunto de nodos a incluir, lo que ayudará a filtrar los resultados finales y a mejorar el algoritmo de búsqueda realizando una búsqueda más enfocada.

Basado en estas condiciones, en caso de que el camino excediera la profundidad definida, si un nodo ya fue visitado (lo que indica la presencia de ciclos) o si el siguiente nodo es parte de las restricciones definidas por el usuario, el algoritmo descartaría el camino y continuaría el análisis. De manera más concreta, en la Figura 23 se muestra el pseudocódigo para la variación de la búsqueda por anchura.

```

function buscarTodosLosCaminosPorAnchura(nodoInicio, nodoFinal,
                                          profundidad, restricciones) {
  Crear lista caminosPosibles
  Crear cola caminosEncontrados
  Agregar nodoInicio a caminosPosibles

  while(caminosPosibles.longitud() > 0) {

    camino = caminosPosibles.primerCamino();

    if(camino.longitud() >= profundidad)
      continue;

    sucesores = obtenerSucesores(camino.ultimoNodo());

    foreach(sucesor : sucesores) {

      boolean esCaminoValido = (sucesor.siguiente() != null &&
                              !camino.contiene(sucesor) &&
                              !restricciones.contiene(sucesor))

      if(esCaminoValido)
        Agregar camino+sucesor a caminosPosibles;

      else if(sucesor == nodoFinal)
        Agregar camino+sucesor a caminosEncontrados;
    }

    return caminosEncontrados
  }
}

```

**Figura 23. Pseudocódigo de búsqueda por anchura para todos los caminos<sup>20</sup>**

Mediante este algoritmo, se encuentran nuevas rutas metabólicas que nacen a partir de complementar los diferentes mapas de entrada. Una vez seleccionada la ruta a utilizar, se pueden filtrar los elementos de modo que se obtengan las proteínas necesarias para continuar el siguiente paso.

---

<sup>20</sup> Figura 23. Elaboración propia.



Con el fin de simplificar las funcionalidades de la aplicación y aprovechar los algoritmos desarrollados, se definieron dos modos a seleccionar por el usuario:

1. Modo básico: busca el camino más corto sin agregar ninguna restricción, inclusión o límite de profundidad. Este permitirá ver visualizaciones de forma rápida y pasar a la siguiente etapa sin agregar configuraciones adicionales al grafo creado.
2. Modo personalizado: calcula todos los caminos posibles, con un límite de profundidad. Además, permite la inclusión o exclusión de compuestos, reacciones o proteínas.

A continuación se muestra un ejemplo del modo básico para encontrar el camino más corto. A partir de un subconjunto de rutas metabólicas, y los puntos de entrada y salida se obtiene el mejor camino bajo las condiciones definidas. En particular, para esta modalidad, se utiliza el algoritmo de Dijkstra y se busca el camino más corto entre dos nodos, sin ninguna restricción, basados solamente en la dirección. Con el fin de simplificar los ejemplos en esta sección, se utilizó el siguiente subconjunto de rutas:

- Metabolismo del Carbono (Apéndice 3)
- Metabolismo de la Galactosa (Apéndice 4)
- Ruta de la Pentosa Fosfato (Apéndice 5)

La Tabla 2 muestra la ruta más corta entre la Lactosa y el Piruvato, para este subconjunto de rutas metabólicas.

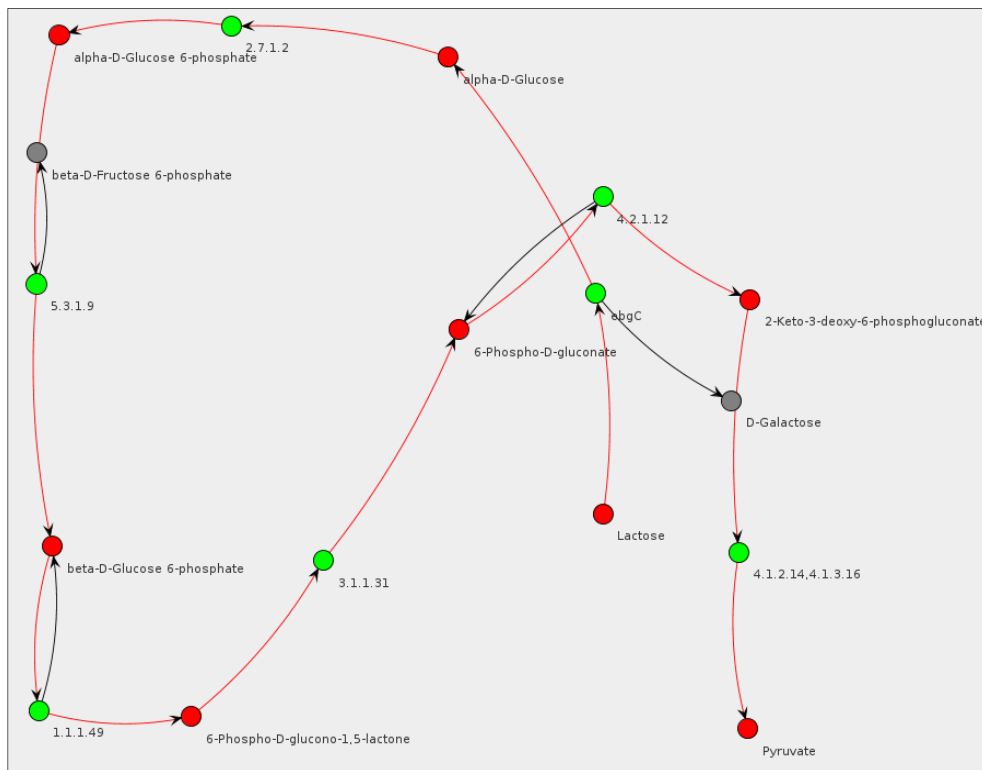
**Tabla 2. Ruta metabólica entre la Lactosa y el Piruvato<sup>21</sup>**

<b>Nodo</b>	<b>Id KEGG</b>	<b>Nombre</b>	<b>Tipo</b>
<b>Inicio</b>	cpd:C00243	Lactose	Compuesto
<b>1</b>	eco:b0344, eco:b3076, eco:b3077	ebgC	Proteína
<b>2</b>	cpd:C00267	alpha-D-Glucose	Compuesto
<b>3</b>	eco:b2388	2.7.1.2	Proteína
<b>4</b>	cpd:C00668	alpha-D-Glucose 6-phosphate	Compuesto
<b>5</b>	eco:b4025	5.3.1.9	Proteína
<b>6</b>	cpd:C01172	beta-D-Glucose 6-phosphate	Compuesto
<b>7</b>	eco:b1852	1.1.1.49	Proteína
<b>8</b>	cpd:C01236	6-Phospho-D-glucono-1,5-lactone	Compuesto
<b>9</b>	eco:b0767	3.1.1.31	Proteína
<b>10</b>	cpd:C00345	6-Phospho-D-gluconate	Compuesto
<b>11</b>	eco:b1851	4.2.1.12	Proteína
<b>12</b>	cpd:C04442	2-Keto-3-deoxy-6-phosphogluconate	Compuesto
<b>13</b>	eco:b1850	4.1.2.14,4.1.3.16	Proteína
<b>Final</b>	cpd:C00022	Pyruvate	Compuesto

El objetivo del algoritmo no es sólo encontrar las rutas entre dos puntos, sino también encontrar las proteínas que se encuentran involucradas en el proceso. Por otro lado, también es importante resaltar todos los compuestos obtenidos a partir de la descomposición de un compuesto más complejo, producto de las reacciones que catalizan las proteínas. A pesar de que estos compuestos no son parte de la ruta crítica, es relevante para el usuario conocer todos los

<sup>21</sup> Tabla 2. Elaboración propia.

subproductos que se producen en una reacción. Estos compuestos se identificarán en las representaciones gráficas por medio de un sistema de etiquetas de color. Las proteínas se resaltarán en verde; los compuestos que se derivan pero que no son parte del camino en gris y la ruta crítica en rojo, como se muestra en la Figura 24.



**Figura 24. Visualización del camino más corto entre la lactosa y el piruvato<sup>22</sup>**

Adicionalmente, el modo personalizado permite la modificación de parámetros adicionales como los siguientes:

<sup>22</sup> Figura 24. Elaboración propia.

- Profundidad del análisis: establece el límite con el cual el algoritmo deja de expandir los nodos del grafo.
- Lista de exclusiones: lista de elementos que no serán tomados en cuenta en los caminos encontrados. Esto permitirá detener la búsqueda en un nodo terminal o filtrar los resultados que contienen este compuesto.
- Lista de inclusiones: lista de elementos que deben ser incluidos obligatoriamente en los resultados, lo que disminuye también la lista final de opciones.

Desde el punto de vista del usuario, la aplicación se limitará a solicitar el compuesto de entrada, el compuesto de salida, la profundidad en la que se buscarán los resultados y los archivos de exclusiones e inclusiones sugeridas por el usuario, como se muestra en la Figura 25.

```

[USER] Please select a start compound. Use default [cpd:C00243] - Lactose? (y/n)
y
[USER] Please select a end compound. Use default [cpd:C00022] - Pyruvate? (y/n)
y
      == Operation Mode ==
      1. [Basic mode] Shortest Path
      2. [Custom mode] All available paths, with restrictions
[USER] Please choose the operation mode [1 or 2]. Default: 2
2
[USER] Please select at which depth you would like to search, maximum depth defined is: 20. Use default
[11]? (y/n)
13
[USER] Do you want to exclude any compound(s)? (y/n)
y
[USER] Please provide an exclusion file, with one compound per line:
exclusions.txt
[USER] Do you want to restrict your paths to any compound(s)? (y/n)
y
[USER] Please provide a valid file, with one compound per line.
inclusions.txt

```

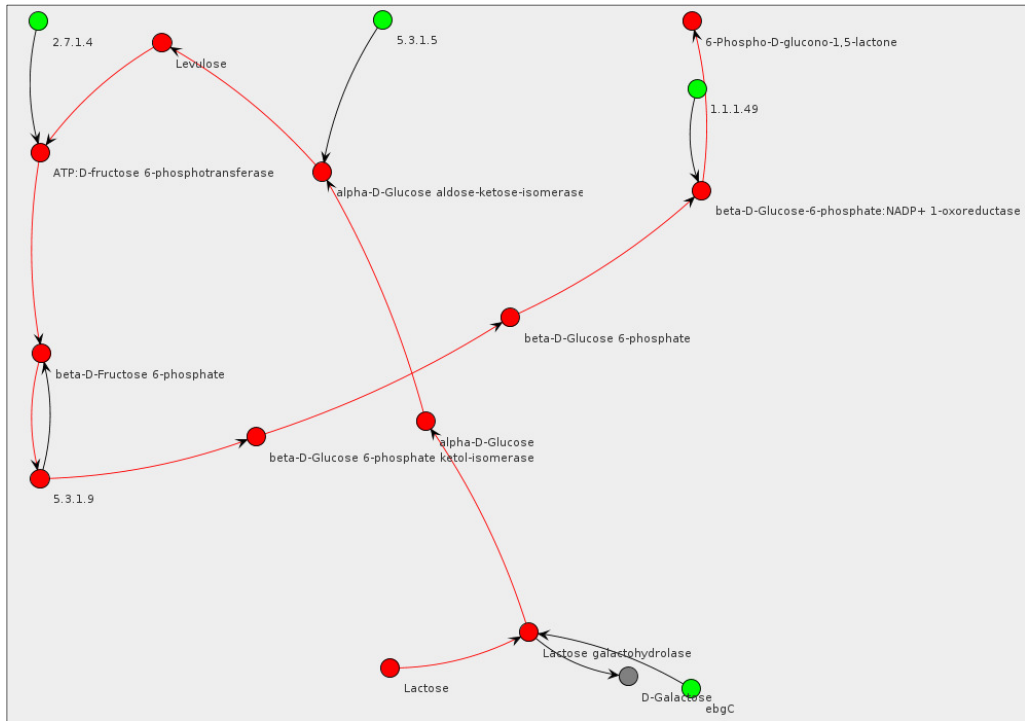
**Figura 25. Selección de entradas para la búsqueda de caminos en modo personalizado<sup>23</sup>**

Con respecto a la estructura de los resultados, debido a la naturaleza de los datos y la forma en que se encuentran conectados, el camino encontrado estará representado en cadenas de la forma: compuesto→ reacción→ compuesto o compuesto → proteína → compuesto. Los caminos poseen proteínas que se encuentran directamente involucrada con los compuestos pero también podría estar conformado por la reacción que cataliza. Existen casos en el que el camino muestra solamente la reacción, pero siempre existe un vínculo con las proteínas que la produce, lo que se interpreta como "La proteína A cataliza la reacción B". Sin embargo, al no ser parte de la ruta principal, no se

---

<sup>23</sup> Figura 25. Elaboración propia.

mostraba en los resultados y se visualizaban solamente como nodos externos a la ruta, como se muestra en Figura 26.



**Figura 26. Visualización del camino incluyendo reacciones y proteínas.<sup>24</sup>**

Como solución y con el fin de hacer un despliegue más claro de la información, una vez que se encontró un subconjunto de caminos, se sustituye la reacción por la proteína que la cataliza. De esta forma, los caminos posibles entre dos compuestos estarían estandarizados en un solo formato compuesto  $\rightarrow$  proteína, como se observó en la Figura 24. Sin importar la modalidad que se elija, se seleccionará un resultado del cual se extraerán las proteínas necesarias para la siguiente etapa.

<sup>24</sup> Figura 26. Elaboración propia.

En la siguiente sección se utilizan los resultados encontrados para el modo básico, con el fin de extraer las proteínas que llevan a cabo el resultado seleccionado. A continuación, se muestran las proteínas que se utilizarán en la siguiente etapa:

- ebgC
- 2.7.1.2
- 5.3.1.9
- 1.1.1.49
- 3.1.1.31
- 4.2.1.12
- 4.1.2.14,4.1.3.16

#### **4.4. Selección de Proteínas**

El siguiente paso en el desarrollo de este prototipo fue la selección de proteínas. A partir de las rutas metabólicas encontradas, se pueden extraer las proteínas capaces de transformar los compuestos de entrada hasta alcanzar el subproducto meta. Para este paso, el usuario seleccionará una ruta, como se muestra en la Figura 27.

```

[INFO] This is the selected path:
cpd:C00243 (Lactose)
  { eco00052 }

eco:b0344,eco:b3076,eco:b3077 (ebgC)
  { eco00052 }

cpd:C00267 (alpha-D-Glucose)
  { eco00052 }

eco:b2388 (2.7.1.2)
  { eco00052 eco01200 }

cpd:C00668 (alpha-D-Glucose 6-phosphate)
  { eco00052 eco00030 }

eco:b4025 (5.3.1.9)
  { eco01200 eco00030 }

cpd:C01172 (beta-D-Glucose 6-phosphate)
  { eco00030 }

eco:b1852 (1.1.1.49)
  { eco01200 eco00030 }

cpd:C01236 (6-Phospho-D-glucono-1,5-lactone)
  { eco01200 eco00030 }

eco:b0767 (3.1.1.31)
  { eco01200 eco00030 }

cpd:C00345 (6-Phospho-D-gluconate)
  { eco01200 eco00030 }

eco:b1851 (4.2.1.12)
  { eco01200 eco00030 }

cpd:C04442 (2-Keto-3-deoxy-6-phosphogluconate)
  { eco01200 eco00030 }

eco:b1850 (4.1.2.14,4.1.3.16)
  { eco01200 eco00030 }

cpd:C00022 (Pyruvate)
  { eco01200 eco00030 }

```

**Figura 27. Selección de la ruta metabólica de interés<sup>25</sup>**

Para extraer los datos de cada proteína se seleccionó UniProt (The Universal Protein Resource, por sus siglas en inglés), un repositorio de datos de secuencias de proteínas e información funcional (UniProt Consortium, 2014). Una de sus ventajas, es que provee una interfaz de programación que permite la extracción de datos por medio de un

---

<sup>25</sup> Figura 27. Elaboración propia.



API REST, ya sea en formato tabulado para ciertas columnas o todas las entradas disponibles en un archivo de texto plano.

Por otro lado, las proteínas encontradas en las rutas metabólicas almacenadas en KEGG, tienen asociado un identificador, denominado "uniprotIds", que permite obtener más datos acerca de ellas. Mediante estas palabras claves es posible extraer información sobre los genes que las producen, los organismos al que pertenecen, su origen y principalmente las referencias cruzadas con otras bases de datos de genes.

La Tabla 3 muestra los identificadores para cada proteína y su respectiva llave en UniProt para poder extraer los campos necesarios.

**Tabla 3. UniProt Ids para las proteínas seleccionadas<sup>26</sup>**

<b>Id KEGG</b>	<b>Nombre</b>	<b>UNIPROT Id</b>
<b>eco:b0344+</b> <b>eco:b3076+eco:b3077</b>	ebgC	P00722 G0ZKW2 P06864 P0AC73
<b>eco:b2388</b>	2.7.1.2	P0A6V8
<b>eco:b4025</b>	5.3.1.9	P0A6T1
<b>eco:b1852</b>	1.1.1.49	P0AC53
<b>eco:b0767</b>	3.1.1.31	P52697
<b>eco:b1851</b>	4.2.1.12	P0ADF6
<b>eco:b1850</b>	4.1.2.14,4.1.3.16	P0A955

<sup>26</sup> Tabla 3. Elaboración propia.

Una vez recolectada esta información, se implementó una conexión basada en Java, previamente provista por la documentación del API (UniProt Consortium, 2012) que permite la extracción de datos en un formato tabulado. A continuación, se describen los campos para el identificador P0A955 como ejemplo:

- Genes: nombre o nombres de los genes que codifican para la secuencia de la proteína de entrada. Ejemplo: eda.
- Secuencia: despliega la secuencia canónica de la proteína. Ejemplo: MKNWK TSAESILTTGPVVP.
- Proteínas: lista exhaustiva de todos los nombres de las proteínas. Ejemplo: KHG/KDPG aldolase; 4-hydroxy-2-oxoglutarate aldolase;2-keto-4-hydroxyglutarate aldolase.
- Nombre de la entrada: identificador mnemónico para la entrada en UniprotKB. Ejemplo: ALKH\_ECOLI.
- Dominio: establece el linaje taxonómico. Ejemplo: Bacteria
- Longitud: largo de la secuencia. Ejemplo: 213.
- Organismo: nombre del organismo fuente de la secuencia de la proteína. Ejemplo: Escherichia Coli (strain K12).
- Verificada: establece si la secuencia fue manualmente anotada y no generada automáticamente. Ejemplo: reviewed.
- Base de datos: referencias cruzadas a entradas relacionadas y encontradas en otras colecciones de datos fuera de UniProtKB,

como GenBank, una base de datos de secuencias genéticas. Ejemplo: X68871, M87458, L20897, X63694, U00096, AP009048.

- Reacción: describe la reacción química que dicha enzima cataliza. Ejemplo: 2-dehydro-3-deoxy-6-phosphate-D-gluconate = pyruvate + D-glyceraldehyde 3-phosphate.
- Función: anotaciones generales o comentarios con respecto a su función. Ejemplo: "Involved in the degradation of glucose via the Entner-Doudoroff pathway. Catalyzes the reversible, stereospecific retro-aldol cleavage of 2-Keto-3-deoxy-6-phosphogluconate (KDPG) to pyruvate and D-glyceraldehyde-3-phosphate. In the synthetic direction, it catalyzes the addition of pyruvate to electrophilic aldehydes with si-facial selectivity. It accepts some nucleophiles other than pyruvate, including 2-oxobutanoate, phenylpyruvate, and fluorobutanoate. It has a preference for the S-configuration at C2 of the electrophile".

En resumen, el proceso de selección de proteínas consiste en: extracción de UniprotIDs de las rutas metabólicas en KEGG, consulta de la información en la base de datos UniProt por medio de la interfaz de programación y descarga de la información a partir del identificador, como se muestra en la Figura 28.



**Figura 28. Proceso de extracción de información de las proteínas<sup>27</sup>**

#### **4.5. Selección de Genes**

En la naturaleza, las proteínas se encuentran codificadas en los distintos genes de uno o varios organismos. Una misma proteína podría ser codificada por diferentes genes, variando ligeramente en su resultado final. En la etapa anterior, se extrajeron diferentes datos de las proteínas entre ellas, las referencias cruzadas que permiten encontrar información adicional de los genes en otras bases de datos.

Para llevar a cabo esta tarea, se seleccionó como base de datos externa a GenBank (Benson, Cavanaugh, & Clark, 2012), una base de datos de secuencias genéticas que posee una colección anotada de todas las secuencias de ADN públicamente disponibles.

A través de este repositorio de datos, es posible extraer información adicional de cada gen, al igual que se hizo en la etapa anterior con las proteínas. Para la construcción de biobricks, la cual será descrita en la etapa siguiente, es necesaria la región codificante como se explica en la sección 3.2.3. Esta es extraída de GenBank, por medio de los

---

<sup>27</sup> Figura 28. Elaboración propia.

identificadores presentes en las referencias cruzadas, como se muestra en la Tabla 4.

**Tabla 4. Referencias cruzadas en GenBank para cada proteína<sup>28</sup>**

<b>KEGG Id</b>	<b>UniProt Id</b>	<b>GenBank Id</b>	
<b>eco:b0344+</b> <b>eco:b3076+eco:b3077</b>	P00722 G0ZKW2 P06864 P0AC73	J01636 V00296 U73857	U00096 AP009048 V00295
<b>eco:b2388</b>	P0A6V8	U22490 U00096 AP009048	
<b>eco:b4025</b>	P0A6T1	X15196 U00006	U00096 AP009048
<b>eco:b1852</b>	P0AC53	M55005 U00096 AP009048 U13783 U13784 U13785 U13786 U13787	U13788 U13789 U13790 U13791 U13792 U13793 U13794 X63694
<b>eco:b0767</b>	P52697	U27192 U00096 AP009048	
<b>eco:b1851</b>	P0ADF6	M87458 X63694 U00096	AP009048 L20897
<b>eco:b1850</b>	P0A955	X68871 M87458 L20897	X63694 U00096 AP009048

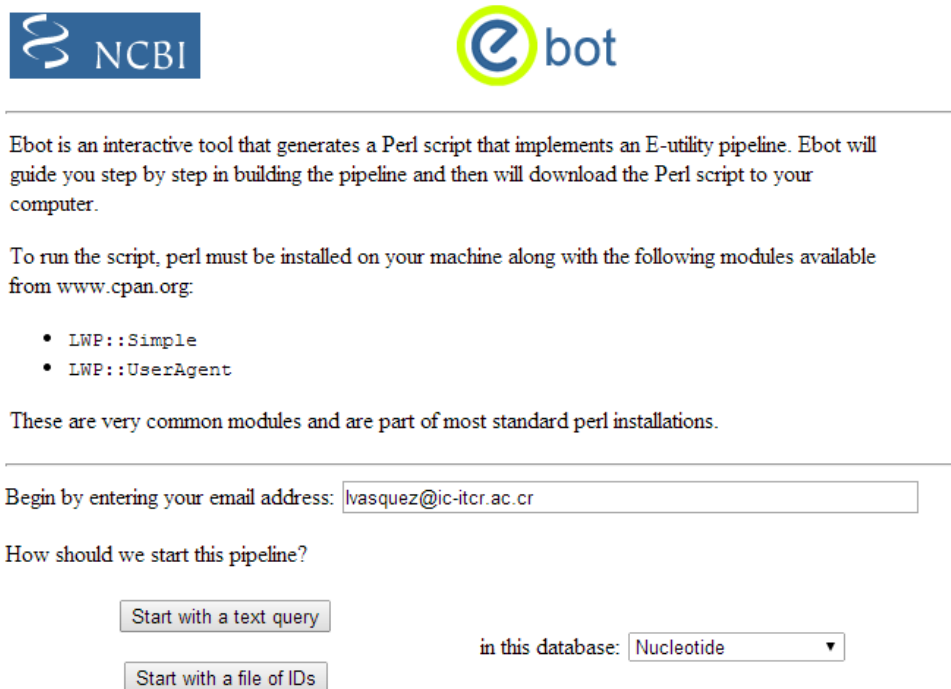
Para establecer una comunicación con GenBank, existe una interfaz de programación provista por "Entrez Programming Utilities" (National Center for Biotechnology Information, 2010), los cuales son un

<sup>28</sup> Tabla 4. Elaboración propia.

conjunto de programas que proveen una interfaz estable a la base de datos del Centro Nacional para la Información Biotecnológica (NCBI, por sus siglas en inglés), del cual GenBank es parte.

Entre estos programas, "Ebot" es una herramienta web interactiva que permite a los usuarios la generación de scripts a partir de una secuencia de pasos personalizados por el usuario (National Center for Biotechnology Information, 2013)

En la Figura 29 se muestra la interfaz web disponible para comenzar configuración:



The image shows the Ebot web interface. At the top left is the NCBI logo, and at the top right is the Ebot logo. Below the logos, there is a horizontal line. The text below the line reads: "Ebot is an interactive tool that generates a Perl script that implements an E-utility pipeline. Ebot will guide you step by step in building the pipeline and then will download the Perl script to your computer." Below this, it says: "To run the script, perl must be installed on your machine along with the following modules available from www.cpan.org:" followed by a bulleted list: "• LWP::Simple" and "• LWP::UserAgent". Below the list, it says: "These are very common modules and are part of most standard perl installations." Below this, there is another horizontal line. The text below the line reads: "Begin by entering your email address:" followed by a text input field containing "lvasquez@ic-itcr.ac.cr". Below the input field, it says: "How should we start this pipeline?" followed by two buttons: "Start with a text query" and "Start with a file of IDs". To the right of the buttons, it says: "in this database:" followed by a dropdown menu showing "Nucleotide".

**Figura 29. Ebot Home page (Sayers, 2010)**

Para poder descargar el CDS a partir del identificador, se debe agregar estos pasos al flujo o "pipeline", como se indica en el siguiente procedimiento:

1. Ingresar el correo electrónico.
2. Seleccione "Start with a text query", con la siguiente base de datos: "Nucleotide".
3. Seleccione "Retrieve records for all dates " y luego "Add Step to Pipeline".
4. Seleccione "Download full records for the data (efetch)" y luego "Build Step".
5. Seleccione el nombre para el archivo de salida en "Output file name" y el formato de salida en "Choose an output format", seleccione: "CDS nucleotide FASTA".
6. Seleccione "End Pipeline".
7. Seleccione el nombre del script de salida.

1	2
ESearch in nuccore with	EFetch CDS nucleotide FASTA to cfs_nucleotide_fasta.out

All done!

Enter the filename for your Perl script:

After clicking the **Generate Perl!** button, please save the script to your computer and then run it by typing **perl** followed by the name of the script in the following window:

- Windows - Command Prompt
- Mac OS X - Terminal
- LINUX/UNIX - Shell

The bulk of the file produced will be routines from the NCBI\_PowerScripting module. The part of the script unique to your pipeline will be in the MAIN BODY of the script starting at line 81:

```

*** SCRIPT MAIN BODY *****
Your pipeline code goes here...
*** END SCRIPT MAIN BODY *****

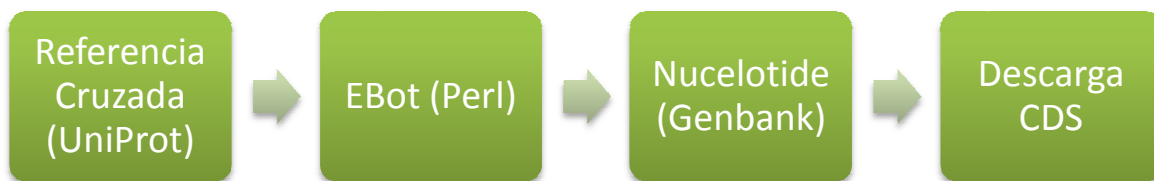
```

### Figura 30. Generación del script para la descarga del CDS por medio de Ebot (Sayers, 2010)

La Figura 30 muestra la configuración seleccionada para el programa el cual está listo para hacer ejecutado por el usuario.

Internamente, el programa se encarga de realizar una búsqueda en la base de datos de Nucléotidos "nuccore" a partir del identificador obtenido en UniProt. Esta base de datos posee una colección de secuencias (genomas y genes) de varias fuentes, incluyendo GenBank. Con la información recopilada, se descarga el CDS correspondiente en un archivo de texto en formato FASTA. El flujo completo de describe en la Figura 31.





**Figura 31. Extracción de la región codificante del gen<sup>29</sup>**

Para el usuario, este proceso se encuentra automatizado, ya que una vez seleccionada las rutas con la cual se va trabajar, la aplicación recopila la información necesaria para la creación del catálogo de biobricks. Es decir, extrae la proteínas necesarias, recopila todas las referencias a GenBank y carga el CDS correspondiente relativo a cada gen, como se muestra en la Figura 32.

```

[INFO] Total of proteins in current path: 7
[INFO] Searching the following query: [P00722, G0ZKW2, P06864, P0AC73, P0A6V8, P0A6T1, P0AC53, P52697, P0ADF6, P0A955]..
[INFO] Connecting to UNIPROT services..
[INFO] Loaded cross-reference into gene for: J01636
[INFO] Loaded cross-reference into gene for: V00296
[INFO] Loaded cross-reference into gene for: U73857
[INFO] Loaded cross-reference into gene for: U00096
[INFO] Loaded cross-reference into gene for: AP009048
[INFO] Loaded cross-reference into gene for: V00295
[WARNING] The protein G0ZKW2_ECOLI is not reviewed. It will be removed from the analysis list
[INFO] Loaded cross-reference into gene for: M64441
[INFO] Loaded cross-reference into gene for: X52031
[INFO] Loaded cross-reference into gene for: X03228
[INFO] Loaded cross-reference into gene for: U18997
[INFO] Loaded cross-reference into gene for: U00096
[INFO] Loaded cross-reference into gene for: AP009048
[INFO] Loaded cross-reference into gene for: M64441
[INFO] Loaded cross-reference into gene for: X52031
  
```

**Figura 32. Búsqueda y carga de genes a partir de referencias<sup>30</sup>**

<sup>29</sup> Figura 31. Elaboración propia.

<sup>30</sup> Figura 32. Elaboración propia.

## **4.6. Construcción del catálogo de biobricks**

La construcción del catálogo está compuesto por 3 fuentes de datos, las siguientes son:

- Genes a partir de análisis de rutas metabólicas: al llevar a cabo el análisis de rutas entre compuestos, se pueden extraer los genes participantes en dichos procesos. Luego, se agregarán a la secuencia los extremos necesarios para poder crear un biobrick estandarizado, como se describe en la sección 3.2.3.
- Biobricks existentes: se obtienen del registro estándar de partes biológicas (International Genetically Engineered Machine Foundation, 2003), según las categorías requeridas por el usuario. Estos se obtienen a partir de una lista de identificadores, las cuales serán cargados localmente para inicializar el catálogo. En caso de que los biobricks no se encuentren disponibles, estos serán descargados directamente de iGem. Para este proyecto, se limitará la lista de biobricks a las categorías establecidas en la sección 3.3.
- Secuencias personalizadas: el usuario puede agregar como parte de los ensamblajes secuencias propias al catálogo por medio de archivos de texto, como se muestra en Figura 33.

```
[INFO] Do you want to add a new sequence to the catalog? (y/n)
y

[USER] Please enter the path of the new sequence file:
sequence_A.txt
Adding new gene from file: sequence_A.txt

[USER] Do you want to add another sequence? (y/n)
y

[USER] Please enter the path of the new sequence file:
sequence_B.txt
Adding new gene from file: sequence_B.txt
```

**Figura 33. Piezas del catálogo ingresadas por el usuario<sup>31</sup>**

Una vez seleccionadas las piezas del catálogo, se procederá a un proceso de selección. Para el caso de los biobricks obtenidos directamente de iGem no se cuenta con información estructurada para todos los elementos, lo que limita establecer algún criterio de selección, salvo las categorías ya mencionadas.

Con respecto a los genes extraídos de las rutas metabólicas, se aplicarán los siguientes criterios:

- **Tamaño:** los genes que tengan un tamaño mayor al definido por la configuración, se excluirán, ya que puede que el segmento descargado corresponda a fragmentos de todo el genoma o varios segmentos que no pertenecen a la región codificante.
- **Sitios de restricción o sitios de corte presentes:** otro criterio a tomar en cuenta son los sitios de restricción presentes. Estos no

---

<sup>31</sup> Figura 33. Elaboración propia.

pueden existir en la secuencia original, ya que afectaría la forma en que las piezas son ensambladas. Los sitios de corte sólo pueden estar en los extremos de cada pieza, de lo contrario, la pieza podría dividir o cortar en el lugar equivocado.

En la Figura 34 se muestran los genes que fueron eliminados bajo estos criterios, los cuales se muestran a manera de advertencia al usuario.

```
[WARNING] This CDS X03228 has more than one segments. It won't be included in the catalog. Besides, it has a size of 60
[WARNING] This CDS L20897 has more than one segments. It won't be included in the catalog. Besides, it has a size of 68
[WARNING] This CDS U00006 has more than one segments. It won't be included in the catalog. Besides, it has a size of 2274
[WARNING] This CDS U18997 has more than one segments. It won't be included in the catalog. Besides, it has a size of 5215
[WARNING] This CDS J01636 has more than one segments. It won't be included in the catalog. Besides, it has a size of 92
[WARNING] This CDS U00096 has more than one segments. It won't be included in the catalog. Besides, it has a size of 62566
[WARNING] This CDS X52031 has more than one segments. It won't be included in the catalog. Besides, it has a size of 74
[WARNING] This CDS M64441 has more than one segments. It won't be included in the catalog. Besides, it has a size of 72
[WARNING] This CDS U27192 has more than one segments. It won't be included in the catalog. Besides, it has a size of 102
[WARNING] This CDS V00295 has more than one segments. It won't be included in the catalog. Besides, it has a size of 25
[WARNING] This CDS M87458 has more than one segments. It won't be included in the catalog. Besides, it has a size of 39
[WARNING] This CDS X63694 has more than one segments. It won't be included in the catalog. Besides, it has a size of 48
[WARNING] This CDS AP009048 has more than one segments. It won't be included in the catalog. Besides, it has a size of 63888
[WARNING] This CDS U73857 has more than one segments. It won't be included in the catalog. Besides, it has a size of 1774
```

**Figura 34. Exclusión de regiones codificantes<sup>32</sup>**

Luego de esta etapa, se pone a disposición el catálogo de biobricks para el ensamblaje de nuevos diseños genéticos, mostrado en la Figura 35. Con el fin de simplificar la lista de componentes del catálogo, se limitó la lista de biobricks, seleccionando sólo un subconjunto de las categorías definidas anteriormente. También, se listan en el catálogo los genes encontrados en el análisis y las secuencias agregadas por el usuario.

---

<sup>32</sup> Figura 34. Elaboración propia.

Genes	Length
====	=====
BBa_C0060	161
BBa_C0061	161
BBa_C0070	161
BBa_C0076	161
BBa_C0078	161
BBa_C0083	161
BBa_C0160	161
BBa_C0161	161
BBa_C0170	161
M55005	1473
U13783	892
U13784	892
U13785	892
U13786	892
U13787	892
U13788	892
U13789	892
U13790	892
U13791	892
U13792	892
U13793	892
U13794	892
U22490	966
V00296	3072
X68871	639
sequence_a	4948
sequence_b	4948

**Figura 35. Catálogo para ensamblajes<sup>33</sup>**

Una vez que se construye el catálogo, se pudieron realizar los distintos ensamblajes según las categorías definidas en la metodología en la sección 3.2.4.

En la Figura 36, se muestra un ejemplo bajo la modalidad "*Biobrick + Plásmido*", en donde el usuario crea un ensamblaje con los genes encontrados en la ruta metabólica seleccionada.

---

<sup>33</sup> Figura 35. Elaboración propia.

```

[INFO] Type of biobricks constructions:
      Option 1 -> 2 Biobrick + Biobrick
      Option 2 -> 2 Biobrick + Plasmid
      Option 3 -> 1 Biobrick + Plasmid

[USER] Please choose one of the 3A assembly available [1-3]:
3

[USER] Choose a biobrick or sequence from the list:
U22490

```

**Figura 36. Modalidades de ensamblaje<sup>34</sup>**

Para construir un nuevo ensamblaje en la modalidad "Biobrick + Plásmido", basta con elegir el biobrick del catálogo. Se utiliza por defecto los sitios de restricción EcoRI y PstI para cortar el segmento. La Figura 37, se resaltan los sitios de corte para el biobrick U22490.

```

GAATTCGGGCGCCTTCTAGATGACAAAGTATGCATTAGTCGGTGATGTGGGCGGCACCAACGCACGCTTTGCTCTGTGTGATATTGCCAGTGGTGAAATCT
CGCAGGCTAAGACCTATTACGGGCTTGATTACCCAGCCTCGAAGCGGTCAATTCGCGTTTATCTTGAAGAACATAAGGTCGAGGTGAAAGACGGCTGTATTG
CCATCGCTTGCCCAATTACCGGTGACTGGGTGGCGATGACCAACCATACCTGGGCGTTCTCAATTGCCGAATGAAAAAGAAATCTCGGTTTTAGCCATCTGG
AAATTATTAACGATTTTACCGGTGATCGATGGCGAACCCGATGCTGAAAAAGAGCATCTGATTTCAGTTTGGTGGCGCAGAACCAGTTCGAAGGTAAGCCTA
TTGCGGTTTACGGTGCCGGAACGGGGCTTGGGGTTGCGCATCTGGTCCATGTCGATAAGCGTTGGGTAAGCTTGCCAGGCGAAGGCGGTACGTTGATTTTG
CGCCGAATAGTGAAGAAGAGGCCATTATCCTCGAAATATTGCGTGCGGAAATTTGGTCATGTTTCGGCGGAGGCGTGCCTTTCTGGCCCTGGGCTGGTGAATT
TGTATCGCGCAATTGTGAAAGCTGACAACCGCCTGCCAGAAAAATCTCAAGCCAAAAGATATTACCGAACGCGCGCTGGCTGACAGCTGCACCGATTGCCGCC
GCGCATTGTGCGTGTGTTGCGTCATTATGGGCCGTTTTGGCGGCAATCTGGCGCTCAATCTCGGGACATTTGGCGGCGTGTGTTATTGCGGGCGGTATCGTGC
CGCCTTCTTGAGTTCTTCAAAGGCTCCGGTTTCCGTGCCGATTTGAAGATAAAGGGCGCTTAAAGAATATGTCCATGATATTCCGGTGTATCTCATCG
TCCATGACAATCCGGCCTTCTCGGTTCCGGTGACATTTACGCCAGACCTTAGGTCACATTCTGTAATACTAGTAGCGGCGCTGCAG

```

**Figura 37. Sitios de corte EcoRI y PstI para el biobrick U22490<sup>35</sup>**

El siguiente paso consiste en seleccionar uno de los plásmidos disponibles (pSB1A3, pSB1T3, pSB3K3), los cuales se cortan con las mismas enzimas de restricción. A la hora de utilizar cada una de las piezas, el extremo del biobrick que fue cortado con una enzima, debe

<sup>34</sup> Figura 36. Elaboración propia.

<sup>35</sup> Figura 37. Elaboración propia.

estar en el extremo del plásmido que se cortó con la misma enzima. En la Figura 38 se selecciona el plásmido pSB1A3 y se resaltan sus sitios de corte.

```
TACTAGTAGCGGCCGCTGCGAGTCCGGCAAAAAAGGGCAAGGTGTACCACCCCTGCCCTTTTCTTTAAAACCGAAAAGATTACTTCGCGTTATGCAGGCTTC
CTCGCTCACTGACTCGCTGCGCTCGGTCGTTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGC
AGGAAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCTGCTGGCGTTTTCCACAGGCTCCGCCCTCAGCAGCATCA
CAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTCCCCCTGGAAGCTCCCTCGTGGCTCTCCTGTTCCGAC
CTGCGGCTTACCGGATACCTGTCCGCTTTCTCCCTTCGGGAAGCGTGGCGCTTCTCATAGCTCAGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTTCG
CTCCAAGCTGGGCTGTGTGCAGAAACCCCGTTTCAGCCGACCGCTGCGCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATC
GCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGTACAGAGTCTTGAAGTGGTGGCTAACTACGGCTACACTAGAAG
AACAGTATTTGGTATCTGCGCTGTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAAACAAACCACCGCTGGTAGCGGTGGTTT
TTTTGTTTGCAGCAGCAGATTACGCGCAGAAAAAAGGATCTCAAGAAAGATCCTTTGATCTTTTACGGGGTCTGACGCTCAGTGGAAACGAAAACTCACG
TTAAGGGATTTGGTCATGAGATTATCAAAAAGGATCTTACCTAGATCCTTTAAATAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAAC
TTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTTCGTTTCCATAGTTGCCTGACTCCCGCTCGTGTAGATAAC
TACGATACGGGAGGGCTTACCATCTGGCCCACTGTGCAATGATACCGCGAGACCAGCTCACCAGGCTCCAGATTTATCAGCAATAAACAGCCAGCCGG
AAGGGCCGAGCGCAGAAAGTGGTCTGCACTTTATCCGCTCCATCCAGTCTATTAATTTGCGGGGAAGCTAGAGTAAGTAGTTCCGCAAGTTAATAGTTT
GCGCAACGTTGTTGCCATTGTACAGGCATCGTGGTGTACGCTCGTCTTTGGTATGGCTTCACTCAGCTCCGGTTCCCAACGATCAAGGGCAGTTACATG
ATCCCCATGTTGTGCAAAAAGCGGTTAGCTCCTTCGGTCTCCGATCGTTGTGAGAAGTAAGTTGGCCGAGTGTATCACTCATGGTTATGGCAGCACT
GCATAATTTCTTACTGTATGCCATCCGTAAGATGCTTTTCTGTACTGGTGTAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTG
CTCTTGGCCGGCGTCAATACGGGATAAATACCGCGCACATAGCAGAACTTTAAAAGTGTCTATCATTGGAAAACGTTCTTCGGGGCGAAAACTCAAGGAT
CTTACCCTGTTGAGATCCAGTTCGATATAACCCACTCGTGCACCCAACCTGATCTTCAAGCATCTTTACTTTCAACAGCGTTCTGGGTGAGCAAAAACAGG
AAGGCAAAAATGCCGCAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTTCTCTTTTCAATATTATTGAAGCATTATCAGGGTTATTG
TCTCATGAGCGGATACATATTTGAATGATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCGAAAAGTGCCACCTGACGCTAAGAAACCAT
TATTATCATGACATTAACCTATAAAAAATAGGCGTATCACGAGGCAGAAATTCAGATAAAAAAATCCTTAGCTTTCGCTAAGGATGATTTCTGAAATTCGGC
GCCGCTTCTAGAG
```

**Figura 38. Sitios de corte EcoRI y PstI para el plásmido pSB1A3<sup>36</sup>**

Una vez que ambas piezas han sido cortadas, se ensamblan tomando en cuenta las enzimas que fueron utilizadas para cortar cada extremo. En la Figura 39 se muestran ambas piezas ya ensambladas, removiendo los sitios de corte en el lugar indicado.

<sup>36</sup> Figura 38. Elaboración propia.

AATTCGCGGCCGCTTCTAGATGACAAAGTATGCATTAGTCGGTGATGTGGCGGCACCAACGCACGCTTGTCTGTGTGATATTGCCAGTGGTAAAATCTC  
GCAGGCTAAGACCTATTACAGGCTTATTACCCAGCCTCGAAGCGGTCATTTCGCGTTTATCTTGAAGAACATAAGGTCGAGGTGAAAGACGGCTGTATTGC  
CATCGCTTGCCAAATACCGGTGACTGGGTGGCGATGACCAACCATACCTGGGCGTCTCAATTGCCGAAATGAAAAAGAATCTCGGTTTTAGCCATCTGGA  
AATTATTAACGATTTTACCCTGTATCGATGGCGAACCCGATGCTGAAAAAGAGCATCTGATTCAGTTTGGTGGCGCAGAACCAGGTCGAAGGTAAGCCTAT  
TGCGGTTTACGGTGCCGGAACGGGGCTTGGGGTTGCGCATCTGGTCCATGTGATAAGCGTTGGTAAGCTTGCAGGCGAAGGCGGTCACGTTGATTTTGC  
GCCGAATAGTGAAGAAGAGGCCATTATCCTCGAAATATTGCGTGGGAAATTTGGTCACTGTTTGGCGGAGGCGTGCCTTTTGGCCCTGGGCTGGTGAATTT  
GTATCGCGCAATTGTGAAAGCTGACAACCGCCTGCCAGAAAATCTCAAGCCAAAAGATATTACCGAACGCGCGCTGGCTGACAGCTGCACCGATTGGCCGCG  
CGCATTGTCGCTGTTTGGCTCATTATGGGCGTTTTGGCGCAATCTGGCGCTCAATCTCGGGACATTTGGCGGCGTGTATTGGCGGCGGTATCGTGCC  
GCGCTTCTTGGATTCTTCAAAGGCTCCGGTTTCCGTGCCGATTTGAAGATAAAGGGCGCTTTAAAGAATATGTCCATGATATTCCGGTGTATCTCATCGT  
CCATGACAATCCGGGCTTCTCGGTTCCGGTGCACATTTACGCCAGACCTTAGGTCACATTTCTGTAATACTAGTAGCGGCCGCTGCAGTCCGGCAAAAAAGG  
GCAAGGTGTACCACCCTGCCCTTTTTCTTAAAACCGAAAAGATTACTTCGCGTTATGCAGGCTTCTCGCTCACTGACTCGCTGCCGCTGGTCTGGTCCGGC  
TGCGGCGAGCGGTATCAGCTCACTCAAAGGCGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAGGCCAGCAAAAAGG  
CCAGGAACCTGAAAAAGGCGCGTGTGCTGGCGTTTTCCACAGGCTCCGCCCTGACGAGCATCAAAAAATCGACGCTCAAGTCAAGGTTGGCGAAACC  
CGACAGGACTATAAGATACCAGGCGTTTTCCCTGGAAGCTCCCTCGTGCCTCTCTGTTCCGACCTGCCGCTTACCGGATACCTGTCCGCTTTCTCC  
CTTCGGGAAGCGTGGCGCTTCTCATAGCTCAGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCCGCTCAAGCTGGGCTGTGTGCACGAACCCCGCTTC  
AGCCCGACCGCTGCGCTTATCCGGTAACTATCGTCTTGGTCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCA  
GAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAG  
TTACCTTCGGA AAAAGAGTTGGTAGTCTTGTATCCGGCAAAACCAACCCGCTGGTAGCGGTGGTTTTTTTGGTTTGAAGCAGCAGATTACGCGCAGAAAAA  
AAGGATCTCAAGAAGATCCTTTGATCTTTCTACGGGTCTGACGCTCAGTGGAACGAAAACACTCAGTTAAGGATTTTGGTCATGAGATTATCAAAAAGGA  
TCTTCACTAGATCCTTTAAATTA AAAATGAAGTTTTAAATCAATCTAAAGTATATAGAGTAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGG  
CACCTATCTCAGCGATCTGTCTATTTGCTTCATCCATAGTTGCTGACTCCCGCTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTG  
CTGCAATGATACCGCGAGACCCAGCTCACCGCTCCAGATTTATCAGCAATAAACAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCTGCAACTTTAT  
CCGCTCCATCCAGTCTATTAATTTGTTGGCGGAAGCTAGAGTAAGTAGTTCCGCAAGTTAATAGTTTGCAGCAAGTTGTTGCCATTGCTACAGGCATCGTGG  
TGTACGCTCGTGGTTGGTATGGCTTCAATCAGCTCCGGTCCCAACGATCAAGGGGAGTTACATGATCCCCATGTTGTGCAAAAAGCGGTTAGTCTCT  
TCGGTCTCCGATCGTTGTGAGAAGTAAAGTTGGCCGAGTGTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTATGCCATCCGTAAGAT  
GTTTTCTGTGACTGGTGAAGTACTCAACCAAGTCACTGAGAATAGTGTATGGCGGACCGAGTTGCTCTTGGCCGCGTCAATACGGGATAATACCGCGC  
CACATAGCAGAACTTTAAAAGTGTCTCATATTGAAAAACGTTCTTGGGGCGAAAACCTCAAGGATCTTACCCTGTTGAGATCCAGTTCGATATAACCCA  
CTCGTGACCAACTGATCTTACGATCTTTACTTTACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAGGGAAATAGGGCGA  
CACGGAATGTTGAATACTCATACTCTTCTTTTCAATATTATTGAAGCATTATCAGGGTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGA  
AAAATAACAATAGGGGTTCCGCGCACATTTCCCGAAAAGTCCACCTGACGCTAAGAAACCATTTATCATGACATTAACCTATAAAAATAGGCGTA  
TCACGAGGCAGAAATTTAGATAAAAAAATCCTTAGCTTTCGCTAAGGATGATTTCTGG

**Figura 39. Unión del biobrick U22490 con el plásmido pSB1A3<sup>37</sup>**

Finalmente, una vez que se concluye el ensamblaje meta, es posible realizar las caracterizaciones estructurales y funcionales sobre el producto final. De esta forma, es posible obtener más información sobre las piezas que se utilizaron en el ensamblaje.

Una limitante en esta etapa del proceso es que si se utilizan biobricks de registro estándar o personalizados por el usuario, no se es posible obtener información de su procedencia o relativos a su función. Sin embargo, si se seleccionan piezas que fueron producto del análisis de rutas metabólicas, es posible extraer detalles de sus funciones, proteínas involucradas y sus respectivas reacciones. En la

<sup>37</sup> Figura 39. Elaboración propia.



Figura 40 se muestran las anotaciones encontradas para el ensamblaje de los ejemplos anteriores.

Structural Annotations  
=====

Number of Nucleotides: 3119  
Number of Amino Acids: 1039

Functional Annotations  
=====

Block: U22490  
Protein: GLK\_ECOLI  
Reaction:  $\text{ATP} + \text{D-glucose} = \text{ADP} + \text{D-glucose 6-phosphate}$ .  
Function: Not highly important in E.coli as glucose is transported into the cell by the PTS system already as glucose 6-phosphate.

**Figura 40. Anotaciones funcionales y estructurales<sup>38</sup>**

## 4.7. Visualización de los datos

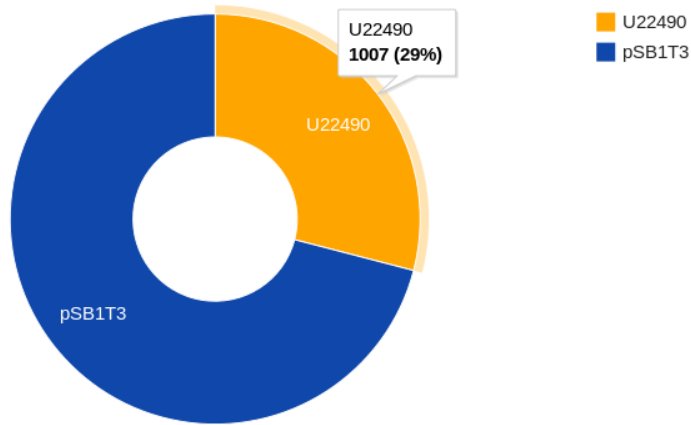
Una vez completado el ensamblaje este se podrá visualizar en un archivo formato HTML, por medio del navegador. Para la implementación de la visualización se utilizó la librería de Google para la creación de gráficos circulares (Google Developers, 2014), la cual se encuentra escrita en Javascript.

En la Figura 41, se visualiza cada pieza, su tamaño y el porcentaje que representa de toda la construcción. Además se desplegará en la parte inferior el nombre de cada elemento, el tipo y la secuencia correspondiente.

---

<sup>38</sup> Figura 40. Elaboración propia.

## DNA Assembly Visualizer



Name	Size	Type	Sequence
U22490	1007	biobrick	GAATTCGGGCGCTTCTAGATGACAAGTATGCATTAGTCGGTATGTGGCGGCACCAAGCGAGCTTCTCTGTGTGATATTGCCAGTGGTGAATCTCGAGGCTAAGACTATTACAGGCTTGAT TACCCAGCCTCGAAGCGTCTATTCGGTTTATCTTGAAGAACAAGTGCAGGTGAAAGACGGCTGTATTGCCATCGCTTGC CCAATTACCGGTGACTGGGTGGCGATGACCAACATACTGGGGCT TCTCAATGCGGAAATGAAAAGAATCTCGTTTATAGCCATCTGGAATATTAAACGATTTTACCGCTGATCGATGGCGAACCCGATGCTGAAAAGAGCACTGATTCAGTTTGGGCGAGAACCGGT CGAAGTAAGCCTATTGCGTTACGTTCCGCGAAGCGGGCTTGGGTTGCGCATCTGGTTCATGTCGATAAGCGTTGGTAAGCTTCCAGCGGAAAGCGGCTACAGTTGATTGCGCCGAATAGT AAGAAGAGGGCATTATCTCGAAATATGCTGCGGAATTTGGTATGTTCCGGAGGCGTGTCTTCTGGCCCTGGTGGTGAATTTGATGCGCAATTTGAAAGCTGACAAACCGCTCCGAGA AAATCTAAGCCAAAAGATATACCGAAGCGGCTGGCTGACAGCTGCACCGATTGCGCGCGGCTGCTGCTGTTTGGCTAATTATGGCGGCTTTTAAAGAAATATGTCATGATATTCGGTATCTCAT TTGGCGGCGTGTATTGCGGGCGGATGTCGCGGCTTCTTGAAGTCTTCAAAGGCTCCGGTTTCCGTCGCGCATTGAAAGATAAGGGCGCTTAAAGAAATATGTCATGATATTCGGTATCTCAT CGTCCATGACAATCCGGGCTTCTGGTTCGGGTGACATTTACGCCAGACTTAGGTACATCTGTAATACTAGTAGCGGCGCTGCGAG

Figura 41. Visualización del ensamblaje<sup>39</sup>

### 4.8. Archivos de salida

Debido a que la mayoría de la interacción con el prototipo se realiza por medio de línea de comandos, se crearon algunos archivos de salida con el fin de proveer más información al usuario. A continuación, se listan brevemente los archivos de salida disponibles:

- nodes.list: contiene la lista de identificadores y nombres compuestos presentes en el grafo. Ejemplo:

```
rn:R00497      gamma-L-glutamyl-L-cysteine:glycine
               ligase (ADP-forming)
cpd:C05577    3,4-Dihydroxymandelaldehyde
eco:b0778,eco:b159 6.3.3.3
```

<sup>39</sup> Figura 41. Elaboración propia.

rn:R00494

glutathione gamma-  
glutamylaminopeptidase

- **paths.list:** muestra todos los caminos encontrados entre los los compuestos dados. Ejemplo:

```
[Path #1] Size: 11
cpd:C00243 (Lactose)
  { eco01100 }
  eco:b0344,eco:b3076,eco:b3077 (ebgC)
  { eco01100 }
cpd:C00267 (alpha-D-Glucose)
  { eco01100 }
  eco:b3565 (5.3.1.5)
  { eco01100 }
cpd:C00095 (Levulose)
  { eco01100 }
  eco:b0394 (2.7.1.4)
  { eco01100 }
cpd:C05345 (beta-D-Fructose 6-phosphate)
  { eco00030 eco01100 }
  eco:b2465,eco:b2935 (2.2.1.1)
  { eco00030 eco01100 }
cpd:C00118 (Glyceraldehyde 3-phosphate)
  { eco00900 eco00030 eco01100 }
  eco:b1850 (4.1.2.14,4.1.3.16)
  { eco00030 eco01100 }
  cpd:C00022 (Pyruvate)
  { eco00900 eco00030 eco01100 }
```

- **viz.html:** Visualización de los ensamblajes en formato HTML.
- **PathwaysAnalyzer.log:** registra los eventos de la aplicación. En caso de estar en modo "debug", imprimirá el contenido de las principales estructuras de datos del sistema.

## 4.9. Tecnologías utilizadas

Para el desarrollo del prototipo, se utilizaron una serie de tecnologías y sistemas, según los requerimientos de cada una de las etapas. En la Tabla 5 se muestran los detalles sobre estas tecnologías.

**Tabla 5. Tecnologías Utilizadas<sup>40</sup>**

Tecnologías	Detalles
<b>Lenguajes</b>	Java <ul style="list-style-type: none"> <li>▪ Versión: 1.7.0_25.</li> <li>▪ Propósito: Implementación del prototipo.</li> </ul> Ruby <ul style="list-style-type: none"> <li>▪ Versión: 1.9.3p0.</li> <li>▪ Propósito: Eliminación de grafos anidados.</li> </ul> Perl <ul style="list-style-type: none"> <li>▪ Versión: v5.14.2.</li> <li>▪ Propósito: Obtener archivos CDS de GenBank.</li> </ul> Javascript <ul style="list-style-type: none"> <li>▪ Goggle javascript API (Google Developers, 2014).</li> <li>▪ Propósito: Visualización de los ensamblajes.</li> </ul>
<b>Editores /Control de Versiones</b>	<ul style="list-style-type: none"> <li>▪ Eclipse Kepler Java EE IDE for Web Developers.</li> <li>▪ VIM - Vi IMproved (version 7.3.429).</li> <li>▪ Eclipse EGit 3.0.3.201309161630-r.</li> </ul>
<b>Sistema Operativo</b>	GNU/Linux <ul style="list-style-type: none"> <li>▪ Xubuntu-12.04.2 Precise Pangolin.</li> <li>▪ Versión del kernel: 3.5.0-45-generic.</li> </ul>
<b>Software de Virtualización</b>	<ul style="list-style-type: none"> <li>▪ Virtualbox 4.3.6 r91406</li> </ul>
<b>Hardware</b>	<ul style="list-style-type: none"> <li>▪ Procesador: Intel(R) Core(TM) i7-3630QM CPU @ 2.40 GHZ (64-bits)</li> <li>▪ RAM: 6 GB</li> </ul>

<sup>40</sup> Tabla 5. Elaboración propia.

## **CAPITULO 5. ANÁLISIS DE RESULTADOS**

El diseño de una herramientas orientada a un fin en particular fue posible por medio de la utilización de datos enfocados y la filtración de resultados de manera controlada. Esto se alcanzó utilizando las rutas metabólicas para una cepa de la bacteria E. Coli y realizando búsquedas sobre compuestos que son relevantes para la descomposición de materias orgánicas, potencialmente utilizadas para la producción de biocombustible. En contraste con otras herramientas de uso genérico, se presenta una propuesta que permite enfocar la investigación a un conjunto de datos que establecerá los análisis hacia un propósito específico.

A pesar de la complejidad de las rutas metabólicas, se confirmó que a través de su análisis es posible la transformación de las mismas en una estructura de datos estándar tales como los grafos. La relevancia de este logro reside en el hecho de poder ubicar un problema del área de la biología en el área de la computación, lo que permite utilizar algoritmos y teorías estudiadas en este campo para el procesamiento y el análisis de datos.

## 5.1. Rutas metabólicas

Para comprobar la hipótesis que se plantea en este trabajo, se realizaron una serie de experimentos que permitieron el diseño de una herramienta que asistiera el proceso de ensamblaje sintético de la bacteria E. Coli.

Inicialmente, se seleccionó una muestra de datos para delimitar el alcance del proyecto y llevar a cabo los experimentos, específicamente con los datos definidos en la sección 3.3.

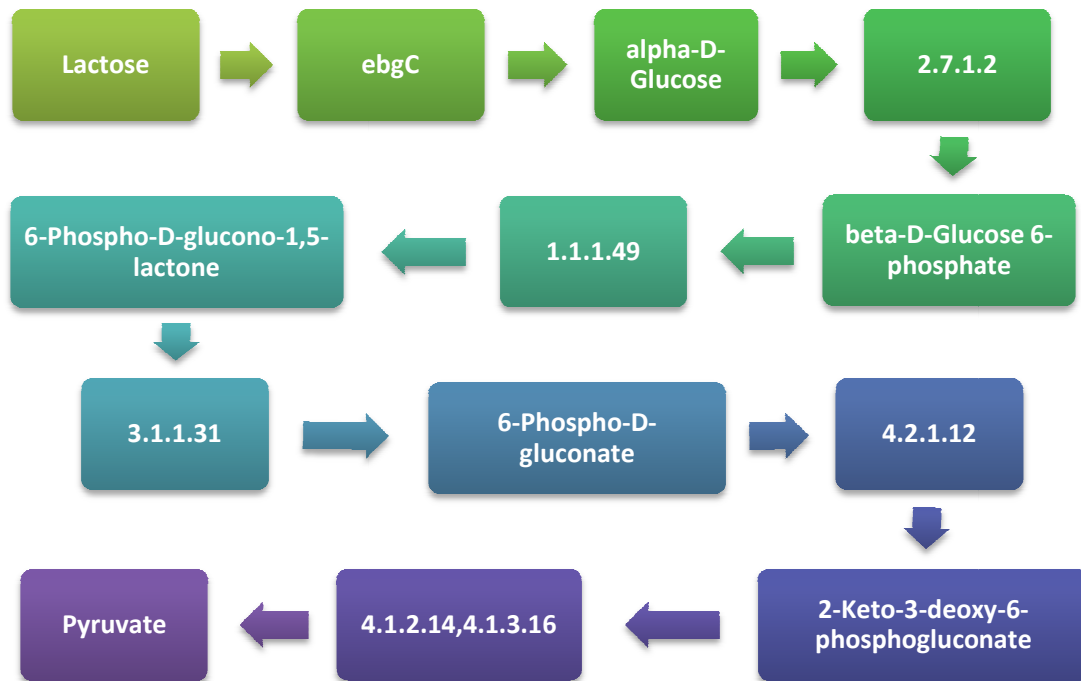
Para seleccionar un conjunto significativo de rutas metabólicas, se realizaron búsquedas en KEGG (Kanehisa Laboratories, 2014) para el organismo Escherichia coli K-12 MG1655. Se recolectaron todos los mapas que coincidieran con las palabras claves "Lactose" y "Pyruvate", como compuestos de entrada y salida, obteniendo un total de 101 rutas metabólicas.

Al unir todas las rutas metabólicas y llevar a cabo el cálculo de rutas hasta una profundidad de 13 nodos, se obtuvieron los siguientes resultados:

- Grafo:
  - 4173 vértices.
  - 5773 aristas.
- Número de caminos encontrados: 14.

- Total de caminos analizados: 229753.

De los 14 posibles caminos entre los dos compuestos, se seleccionó el primero para llevar a cabo los análisis:



**Figura 42. Ruta metabólica entre la Lactosa y el Piruvato<sup>41</sup>**

Cada nodo puede provenir de una misma ruta o existe la alternativa de crear conexiones entre los nodos en común de los distintos mapas. Esto permite que se sugieran nuevos caminos, que no se encuentran explícitos en el organismo analizado. La Tabla 6 muestra los mapas involucrados en la ruta metabólica seleccionada en la Figura 42.

<sup>41</sup> Figura 42. Elaboración propia.

**Tabla 6. Caminos a partir de 101 rutas metabólicas<sup>42</sup>**

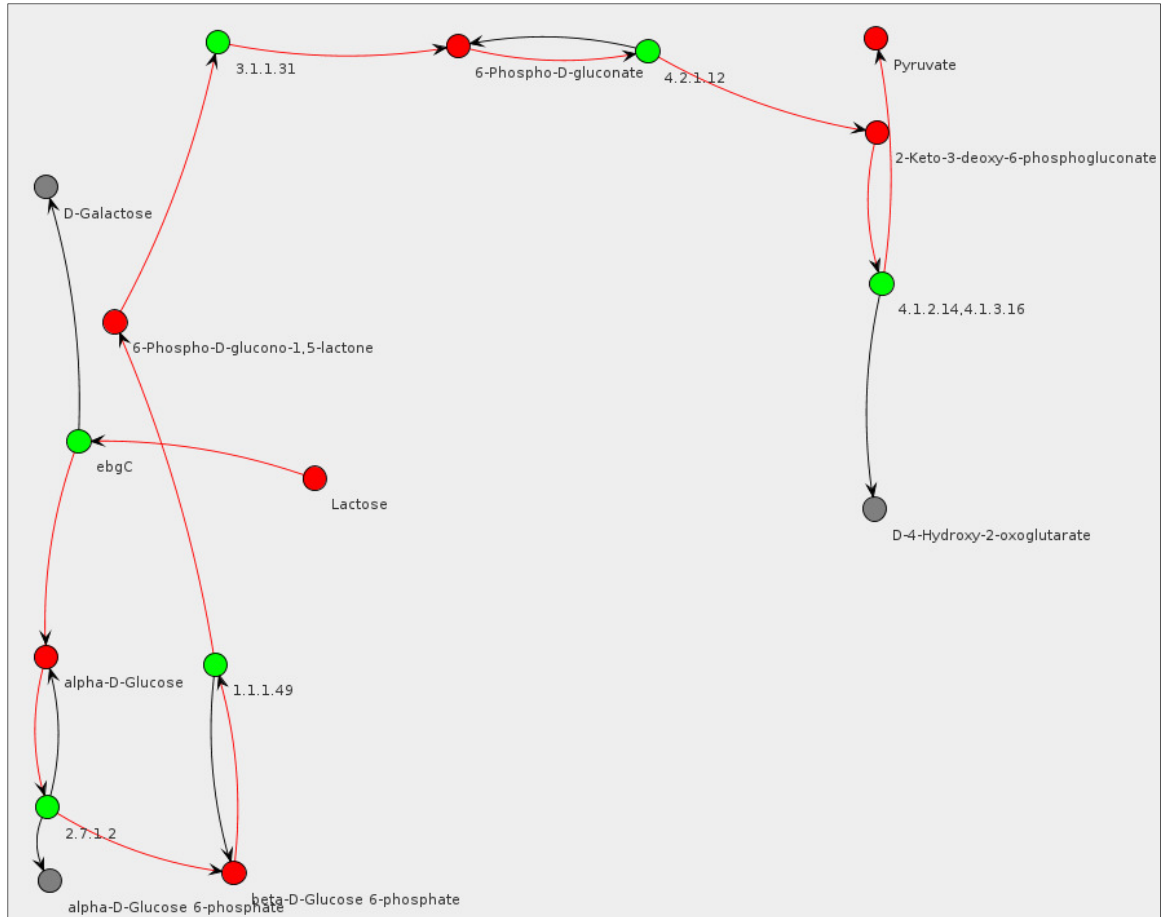
<b>Id</b>	<b>Nombre</b>	<b>Mapas Padres</b>	
<b>cpd:C00243</b>	Lactose	eco00052	
<b>eco:b0344+</b> <b>eco:b3076+</b> <b>eco:b3077</b>	ebgC	eco00052 eco00511	
<b>cpd:C00267</b>	alpha-D-Glucose	eco00010 eco00051 eco00052 eco00500	eco00520 eco01110 eco01120
<b>eco:b2388</b>	2.7.1.2	eco00010 eco00052 eco00500 eco00520	eco00521 eco01110 eco01120 eco01200
<b>cpd:C01172</b>	beta-D-Glucose 6-phosphate	eco00010 eco00030 eco00500	eco01110 eco01120
<b>eco:b1852</b>	1.1.1.49	eco00030 eco01110	eco01120 eco01200
<b>cpd:C01236</b>	6-Phospho-D-glucono-1,5-lactone	eco00030 eco01110	eco01120 eco01200
<b>eco:b0767</b>	3.1.1.31	eco00030 eco01110	eco01120 eco01200
<b>cpd:C00345</b>	6-Phospho-D-gluconate	eco00030 eco01110	eco01120 eco01200
<b>eco:b1851</b>	4.2.1.12	eco00030 eco01120	eco01200
<b>cpd:C04442</b>	2-Keto-3-deoxy-6-phosphogluconate	eco00030 eco01120	eco01200
<b>eco:b1850</b>	4.1.2.14,4.1.3.16	eco00030 eco00330	eco01120 eco01200
<b>cpd:C00022</b>	Pyruvate	eco00010 eco00020 eco00030 eco00053 eco00250 eco00260 eco00270 eco00290 eco00330 eco00362 eco00620 eco00621 eco00622	eco00650 eco00660 eco00680 eco00760 eco00770 eco00900 eco01110 eco01120 eco01200 eco01210 eco01220 eco01230

<sup>42</sup> Tabla 6. Elaboración propia.



Del análisis de los compuestos presentes en los mapas, se determinó que aunque un grupo de compuestos comunes estén presentes en varios mapas, no necesariamente se encuentran conectados, es decir, hay mapas que aportan solamente la conexión entre dos compuestos. Al unir los caminos entre las distintas rutas de entrada, si existen nodos que ya han sido reportados por alguno de los mapas analizados previamente, el camino se podrá construir por un subconjunto simplificado de ellos.

Otro aspecto es que los grafos tienden a aumentar considerablemente de tamaño con algunas rutas metabólicas, por lo que se tiene la limitante de no poder graficar toda la estructura. Sin embargo, dado a que los caminos resultantes tienen un tamaño significativamente más pequeño, pueden visualizarse por medio de las librerías utilizadas, como se muestran en la Figura 43.



**Figura 43. Visualización del camino mostrado en la Tabla 6<sup>43</sup>**

En la base de datos KEGG se manejan nodos de diversos tipos, por ejemplo, compuestos, reacciones, moléculas, proteínas, entre otros. Todos los elementos están altamente conectados, lo que implica que el camino mostrado se podría basar en las reacciones y no en las proteínas que las catalizan.

Con respecto a la validación de la estructura de datos y los caminos encontrados, se llevaron a cabo las siguientes tareas:

<sup>43</sup> Figura 43. Elaboración propia.

- Comparación de resultados con rutas metabólicas originales: debido a que la lista de resultados finales contiene información sobre la procedencia de cada nodo, se pudo identificar el conjunto de mapas que debían ser validados para cada ruta. Para llevar a cabo esta etapa, se seleccionó un subconjunto de caminos y se verificó la conectividad establecida entre cada nodo con respecto a los archivos de entrada. Como guía se utilizaron los mapas originales, descargados directamente del sitio web, los cuales no poseen ningún tipo de transformación o filtro. Esto permitió descartar incongruencias que se pudieran dar durante el procesamiento de los archivos KGML a los grafos o diferencias de conectividad entre los compuestos de la ruta. De esta forma, fue posible validar que la estructura de datos, estableciendo que era una representación equivalente para llevar a cabo los análisis planteados.
- Validación de los usuarios: una vez completada la etapa anterior, se crearon casos de prueba en los cuales se incluyeron conjuntos de rutas posibles entre el Piruvato y la Lactosa, indicando los mapas fuentes utilizados en el análisis. Luego, estos se enviaron a los usuarios que colaboraron en el planteamiento inicial de esta investigación y que además son parte del proyecto para la producción de biocombustibles a partir del suero de la leche, por

medio de Biología Sintética. En particular, el Ing. Ricardo Alvarado del Laboratorio Nacional de Nanotecnología (LANOTEC), área nanobiotecnología y con experiencia en proyectos enfocados en la producción de biocombustibles, realizó una revisión de los resultados, verificando la conectividad entre los nodos, los subproductos producidos por cada reacción y las proteínas involucradas en el proceso. A partir de esto, se brindaron las observaciones pertinentes del caso, que permitieron mejorar la calidad y precisión de los resultados del prototipo.

## **5.2. Proceso de creación del catálogo y los ensamblajes**

A partir de camino anterior, se identifican varias proteínas, las cuales siempre tendrán un prefijo como "eco" es su nombre y "protein" en el atributo de cada nodo relacionado con el tipo.

Estos resultados confirman la factibilidad del cálculo y análisis de rutas por medio de grafos construidos a partir de rutas metabólicas ya preestablecidas. Además, a pesar de las limitaciones a partir de una profundidad 16, al menos para este caso, se presentan suficientes resultados para la meta de las herramientas (más de 28,000 rutas posibles).

Cualquier de estos caminos permite la extracción de proteínas, las cuales poseen las referencias de los genes que la producen. Para el caso anterior, se tienen los siguientes resultados:

- **ebgC (eco:b0344+eco:b3076+eco:b3077)**
  - Uniprot ID: P00722, G0ZKW2, P06864, P0AC73
  - Referencias cruzadas (GenBank):
    - J01636
    - V00296
    - U73857
    - U00096
    - AP009048
    - V00295
  
- **2.7.1.2 (eco:b2388)**
  - Uniprot ID: P0A6V8
  - Referencias cruzadas (GenBank):
    - U22490
    - U00096
    - AP009048
  
- **1.1.1.49 (eco:b1852)**
  - Uniprot ID: P0AC53
  - Referencias cruzadas (GenBank):
    - M55005
    - U00096
    - AP009048
    - U13783
    - U13784
    - U13785
    - U13786
    - U13787
    - U13788
    - U13789
    - U13790
    - U13791
    - U13792
    - U13793
    - U13794
    - X63694
  
- **3.1.1.31 (eco:b0767)**
  - Uniprot ID: P52697
  - Referencias cruzadas (GenBank):

- U27192
- U00096
- AP009048
  
- **4.2.1.12 (eco:b1851)**
  - Uniprot ID: P0ADF6
  - Referencias cruzadas (GenBank):
    - M87458
    - X63694
    - U00096
    - AP009048
    - L20897
  
- **4.1.2.14,4.1.3.16 (eco:b1850)**
  - Uniprot ID: P0A955
  - Referencias cruzadas (GenBank):
    - X68871
    - M87458
    - L20897
    - X63694
    - U00096
    - AP009048

Es importante mencionar que no todas los genes enlistados son aptos para el ensamblaje y la creación de nuevos diseños, debido los criterios ya descritos anteriormente. Sin embargo, se muestra aquellos que cumplen y también los que no son aptos por medio de una advertencia. En la Tabla 7, se detallan cada uno de los casos, para el ejemplo utilizado.

**Tabla 7. Caracterización de genes encontrados<sup>44</sup>**

Referencia	Tamaño	Descripción
<b>V00295</b>	25	Posee varios segmentos, excede el máximo definido.
<b>M87458</b>	39	Posee varios segmentos, excede el máximo definido.
<b>X63694</b>	48	Posee varios segmentos, excede el máximo definido.
<b>X03228</b>	60	Posee varios segmentos, excede el máximo definido.
<b>L20897</b>	68	Posee varios segmentos, excede el máximo definido.
<b>M64441</b>	72	Posee varios segmentos, excede el máximo definido.
<b>X52031</b>	74	Posee varios segmentos, excede el máximo definido.
<b>J01636</b>	92	Posee varios segmentos, excede el máximo definido.
<b>U27192</b>	102	Posee varios segmentos, excede el máximo definido.
<b>X68871</b>	639	Válido
<b>U13783</b>	892	Válido
<b>U13784</b>	892	Válido
<b>U13785</b>	892	Válido
<b>U13786</b>	892	Válido
<b>U13787</b>	892	Válido
<b>U13788</b>	892	Válido
<b>U13789</b>	892	Válido
<b>U13790</b>	892	Válido
<b>U13791</b>	892	Válido
<b>U13792</b>	892	Válido
<b>U13793</b>	892	Válido
<b>U13794</b>	892	Válido
<b>U22490</b>	966	Válido
<b>M55005</b>	1473	Válido
<b>U73857</b>	1774	Posee varios segmentos, excede el máximo definido.
<b>V00296</b>	3072	Válido
<b>U18997</b>	5215	Posee varios segmentos, excede el máximo definido.
<b>U00096</b>	62566	Posee varios segmentos, excede el máximo definido.
<b>AP009048</b>	63888	Posee varios segmentos, excede el máximo definido.

Al solicitar un CDS por medio de una referencia, los resultados pueden contener fragmentos del genoma, por lo que no es posible ensamblarlo como una pieza individual. Como una solución para estos casos, se cuentan la cantidad de fragmentos y el largo de los resultados de la

<sup>44</sup> Tabla 7. Elaboración propia.

referencia. Más adelante, en el momento de los ensamblajes, se determina si también posee algún sitio de restricción que afecte el proceso.

Utilizando un subconjunto de los biobricks del Apéndice 9, se construye un catálogo para llevar a cabo los ensamblajes.

**Tabla 8. Catálogo Completo<sup>45</sup>**

<b>Biobricks</b>	<b>Tamaño</b>
BBa_C0060	161
BBa_C0061	161
BBa_C0070	161
BBa_C0076	161
BBa_C0078	161
BBa_C0083	161
BBa_C0160	161
BBa_C0161	161
BBa_C0170	161
M55005	1473
U13783	892
U13784	892
U13785	892
U13786	892
U13787	892
U13788	892
U13789	892
U13790	892
U13791	892
U13792	892
U13793	892
U13794	892
U22490	966
V00296	3072
X68871	639

<sup>45</sup> Tabla 8. Elaboración propia.



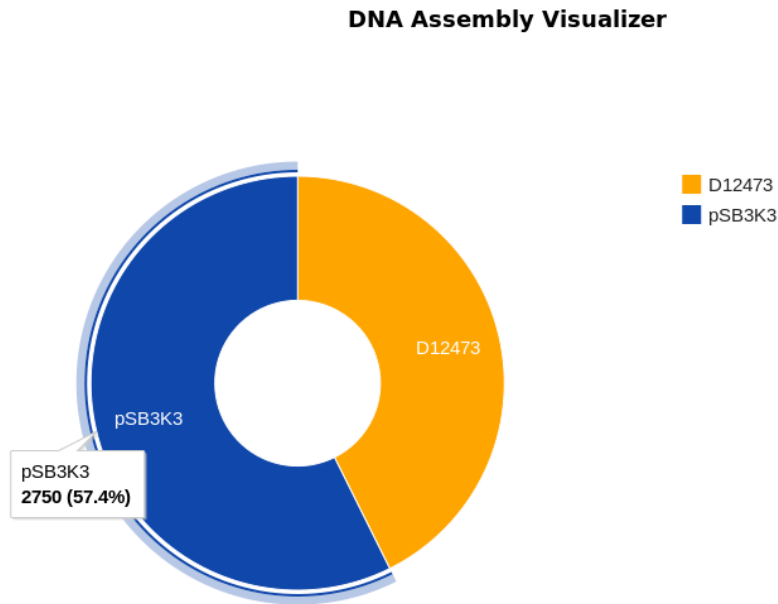
El análisis realizado permitió la creación y validación de un catálogo con los componentes necesarios para los ensamblajes sintéticos de la bacteria E.Coli, específicamente el organismo Escherichia coli K-12 MG1655. Este catálogo permitirá el diseño de nuevos organismos para la producción de biocombustibles, mediante la mejora de sus rutas metabólicas, cumpliendo así con el enunciado del primer objetivo.

Con el fin de validar el enfoque de ensamblaje por medio de estos componentes genéticos, se siguió la metodología definida por iGem, en donde se ensamblan los biobricks en plásmidos o entre sí, con el fin de comprobar su compatibilidad, entre ellos.

Así mismo, el prototipo es capaz de definir 3 tipos de ensamblajes ya mencionados y detallados en la sección 3.2.4.

Con respecto a la caracterización de los ensamblajes, es posible medir estructuralmente su longitud y número de amino ácidos. Por el lado de las anotaciones funcionales, si los componentes provienen de las rutas metabólicas se dispone de información relevante como la reacción que cataliza la enzima y una breve descripción de la función que lleva a cabo. Estos dos enunciados completan el segundo y el tercer objetivo definidos en la investigación.

Finalmente, por medio de computación gráfica y utilizando tecnologías web, es posible mostrar los ensamblajes terminados y algunas de sus características de ellos, como se muestra en la Figura 44.



**Figura 44. Visualizador de ensamblajes<sup>46</sup>**

Con este último resultado, se verifican y se cumplen los objetivos planteados, comprobando que es posible el diseño de una herramienta para la biología sintética enfocada en un dominio en específico como lo es la producción de biocombustibles.

---

<sup>46</sup> Figura 44. Elaboración propia.

### 5.3. Escalabilidad de la solución

Para medir la escalabilidad de la solución se utilizó la totalidad de las rutas metabólicas que coincidieron con los compuestos mencionados. En total se obtuvieron 102 rutas metabólicas, las cuales se detallan en el Apéndice 8. Luego, se realizaron búsquedas hasta una profundidad de 17 nodos sin ninguna restricción. A partir de este escenario, se obtuvieron 28,000 caminos posibles luego de haber analizado un aproximado de 85 millones de nodos, como se muestra en la Tabla 9.

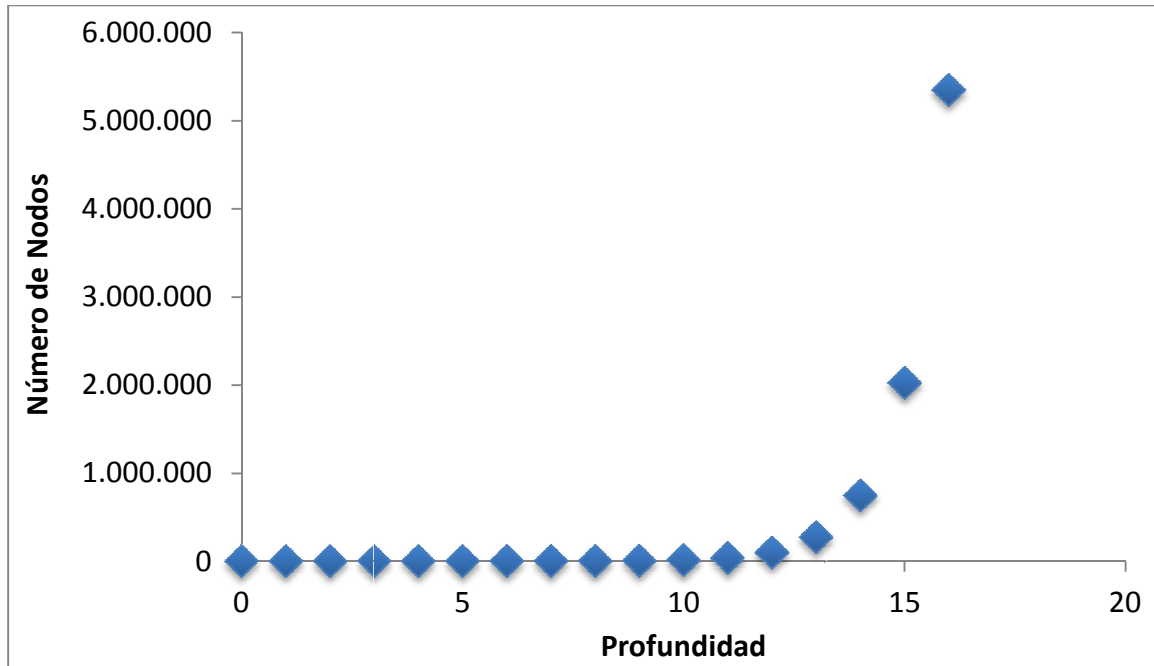
**Tabla 9. Caminos encontrados hasta profundidad 16<sup>47</sup>**

<b>Profundidad</b>	<b>Caminos Analizados</b>	<b>Caminos Encontrados</b>	<b>Total Nodos Analizados</b>
0	0	0	0
1	1	0	1
2	2	0	4
3	4	0	12
4	18	0	72
5	59	0	295
6	189	0	1.134
7	520	0	3.640
8	1.560	0	12.480
9	4.457	0	40.113
10	12.711	0	127.110
11	35.033	28	385.363
12	97.896	66	1.174.752
13	270.494	768	3.516.422
14	744.980	2.372	10.429.720
15	2.022.439	9.566	30.336.585
16	5.343.934	28.512	85.502.944

<sup>47</sup> Tabla 9. Elaboración propia.

Dado a que los experimentos ejecutados no involucraban ninguna restricción por parte de usuario, representaba una búsqueda exhaustiva hasta llegar a la meta o a la profundidad definida, por lo que luego de 17 niveles de profundidad no era posible continuar la búsqueda debido a una limitante de recursos.

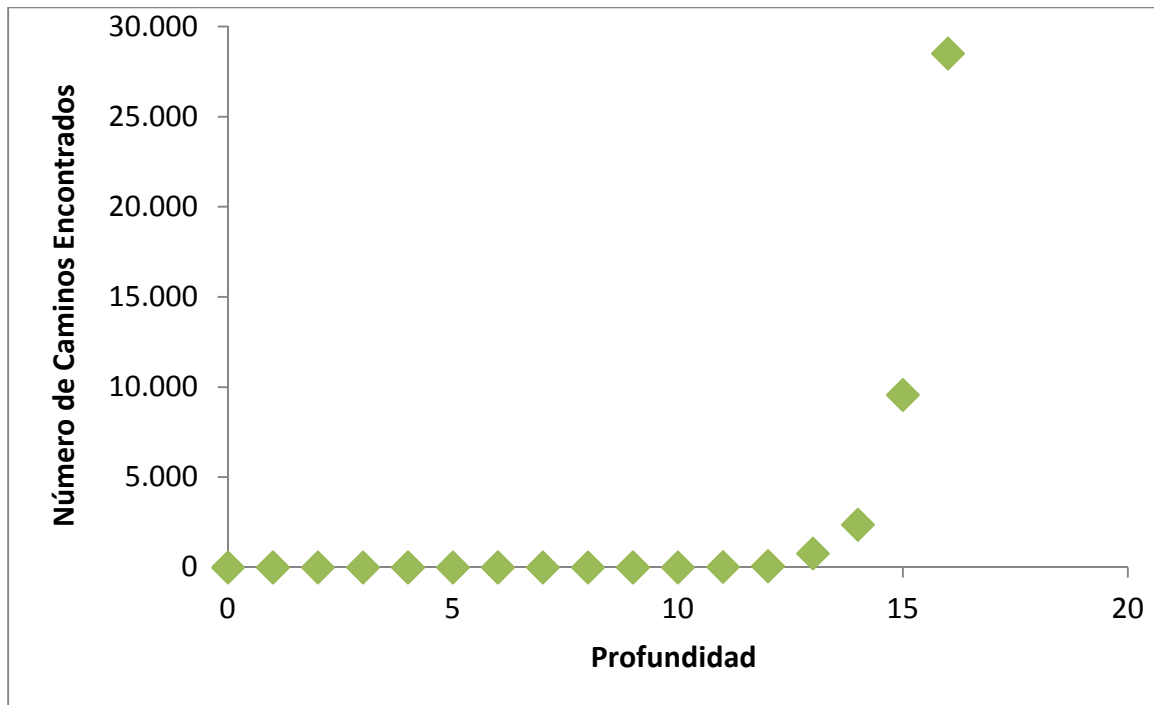
También, debido a la distancia de ambos compuestos, era necesario realizar un análisis hasta 13 niveles de profundidad para encontrar posibles caminos, los cuales van aumentando significativamente en cada iteración. En la Figura 45, se muestra el aumento de nodos por cada nivel que se avanza, expandiendo cada vez más el árbol de soluciones hasta llegar a su límite.



**Figura 45. Número de caminos analizados<sup>48</sup>**

Lo mismo ocurre con el número de soluciones dadas, ya que cada nivel, agrega posibles alternativas a los resultados existentes. Al llegar al límite de capacidad del algoritmo, el número de soluciones es suficiente o incluso más que eso, lo que cumple el objetivo definido.

<sup>48</sup> Figura 45. Elaboración propia.



**Figura 46. Número de caminos encontrados<sup>49</sup>**

Por medio de este análisis, se considera que no es una imitante de alto impacto ya que el número de resultados es suficiente y podría aumentar según las restricciones que el usuario defina.

Las pruebas se corrieron en una máquina virtual, en un sistema GNU/Linux. En la Tabla 10 se detalla el hardware utilizado para la ejecución de las pruebas.

<sup>49</sup> Figura 46. Elaboración propia.

**Tabla 10. Hardware para pruebas de rendimiento<sup>50</sup>**

	<b>Procesador</b>	<b>RAM</b>	<b>Sistema Operativo</b>	<b>Detalles</b>
<b>Sistema Anfitrión</b>	Intel(R) Core(TM) i7-3630QM CPU @ 2.40 GHZ (64-bits)	6,00 GB	Windows 8	N/A
<b>Sistema Huésped</b>	Compartido con sistema anfitrión	3,00 GB	Xubuntu-12.04.2 Precise Pangolin	Virtualbox 4.3.6 r91406

## **5.4. Limitaciones**

Luego de cumplir con los objetivos propuestos, existen limitaciones tanto de los datos de entrada cómo de la herramienta que representan a la vez una oportunidad de mejora para el proyecto. A continuación se detallan aquellas que pertenecen al conjunto de datos con el que se trabajó y también, las que son correspondientes al prototipo desarrollado.

### **5.4.1. Limitaciones del prototipo**

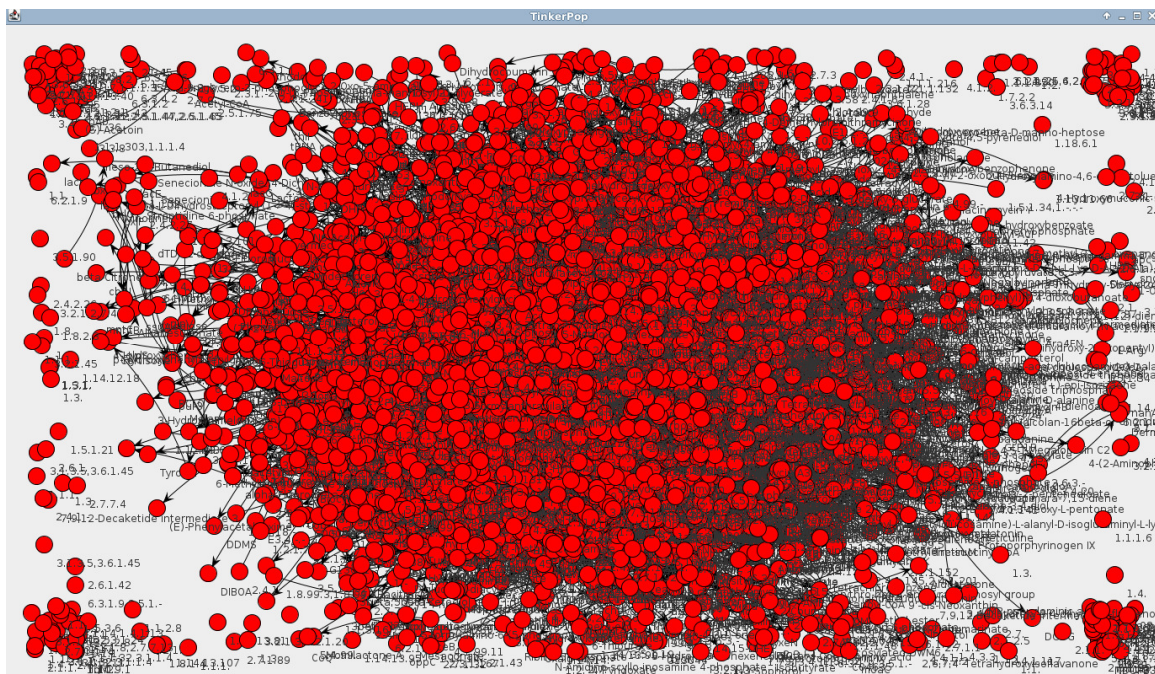
#### *a. Visualización de grafos de tamaño significativo*

Las librerías utilizadas para la implementación de los grafos incluyen módulos que facilitan la visualización de la estructura. Sin embargo, cuando la densidad de elementos es muy alta, no es posible crear una

---

<sup>50</sup> Tabla 10. Elaboración propia.

graficación de la estructura completa que posea una interpretación visualmente coherente como se muestra en la Figura 47.



**Figura 47. Visualización de grafos con alta densidad<sup>51</sup>**

*b. Transformación de rutas dependiente del KEGGtranslator*

Con el fin de traducir las rutas metabólicas a un formato fácil de analizar y navegar se utilizó la herramienta KEGGtranslator. Sin embargo, al llevar experimentos con mapas que pertenecían a organismos de referencia y no uno en particular, la herramienta no podía llevar a cabo la traducción al formato meta a pesar de que si

<sup>51</sup> Figura 47. Elaboración propia.



lograba interpretar y escribir en otros formatos como SBML. Al realizar la traducción al formato GraphML, la herramienta caía en un estado de error, como se muestra en la Figura 48.

```
KEGGtranslator
Copyright © 2010–2012 the University of Tuebingen,
Center for Bioinformatics Tuebingen (ZBIT).
This program comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome
to redistribute it under certain conditions.
See http://www.gnu.org/licenses/lgpl-3.0-standalone.html for details.
-----
Mar 15, 2014 5:39:10 PM de.zbit.Launcher launchCommandLineMode --- INFO:
Launching command-line mode of KEGGtranslator.
Mar 15, 2014 5:39:11 PM de.zbit.kegg.io.AbstractKEGGtranslator
preProcessPathway --- INFO: Fetching information from KEGG online
resources...
Mar 15, 2014 5:39:27 PM de.zbit.kegg.io.AbstractKEGGtranslator
preProcessPathway --- INFO: Information fetched. Translating pathway...
100% | Node 39/39
java.lang.NullPointerException
```

**Figura 48. Error de traducción en KEGGtranslator<sup>52</sup>**

Con respecto al soporte de esta herramienta, no existe un foro público de discusión para errores comunes y se trató de contactar a los autores en busca de soporte sin éxito. Además, no hubo respuesta alguna que explicara detalladamente la causa de dichos problemas. Actualmente, el prototipo utiliza mapas de organismos en específico, lo que dio buenos resultados con el traductor.

### *c. Recursos limitados para el cálculo de rutas*

---

<sup>52</sup> Figura 48. Elaboración propia.

Para la implementación de los algoritmos de búsqueda se trató de utilizar el mínimo de datos y estructuras de datos para el uso óptimo de la memoria. Sin embargo, dado a que los recursos son limitados y se maneja una cantidad considerable de datos, el prototipo mostrará una excepción de memoria "java.lang.OutOfMemoryError". Los experimentos realizados mostraron que la limitante se muestra para cuando se han analizado más de 85,502,944 de nodos. Sin embargo, se determinó que tiene un bajo impacto, ya que para ese momento se tienen al menos 28,000 caminos posibles a analizar, los cuales son un número significativo y hasta excesivo para el análisis individual.

#### *d. Filtración de CDS*

Los CDS de cada gen se obtienen a partir de las referencias externas de las proteínas. Sin embargo, estas referencias no siempre apuntan a esta región sino también a fragmentos del genoma. Al no tener mayor información además de la secuencia, se garantiza que un elemento es un CDS si no es mayor que cierto límite definido y si no posee ningún otro fragmento. Sin embargo, podría existir un CDS mayor al límite definido por ejemplo. Como solución alterna, en caso de que esta metodología excluya algún CDS relevante para el usuario, se propone habilitar el ingreso de nuevas secuencias seleccionadas por el usuario.

## **5.4.2 Limitaciones de los datos de entrada**

*a. Descarga masiva de mapas de KEGG limitada por el alto costo de las licencias*

A partir 2011, KEGG introdujo un sistema de suscripciones para acceder los módulos de descarga masiva por medio de FTP, como respuesta a recortes presupuestarios realizados por la Agencia de Tecnología y Ciencia de Japón (Hayden, 2013). Para fines académicos, la licencia para uso personal tiene un costo de \$2000 anuales y para uso organizacionales son \$5000 anuales, para todos aquellos países que se encuentren fuera de Japón.

Sin embargo, utilizar la información del sitio web de KEGG tiene ningún costo para los usuarios, lo que permite la descarga individual de los mapas a través su interfaz de programación (API), con la limitante que se debe compilar previamente la lista de mapas que se quiere descargar y utilizar.

*b. Arcos entre dos compuestos dependientes de cada ruta metabólica*

Debido a la naturaleza de las rutas metabólicas, las conexiones entre dos nodos no siempre se encuentran presentes para un mismo

subconjunto de nodos. Es decir, si dos compuestos se encuentran presentes en un mapa u otro, no necesariamente existirá una conexión entre ellos.

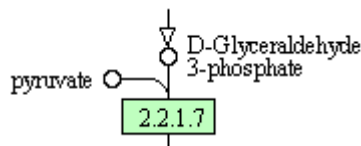
Por ejemplo, utilizando la siguiente muestra de mapas:

- eco00010.
- eco00052.
- eco00051.
- eco00900.
- eco00030.

Y se analiza las conexiones entre:

- D-Glyceraldehyde-3P.
- Pyruvate.

No hay un arco que conecte ambos a pesar de que los compuestos se encuentran presentes en varios de estos mapas. Sin embargo, al agregar el mapa eco01100, se establece una conexión entre ambos, como se muestra en la Figura 49.



**Figura 49. Conexión entre el D-Glyceraldehyde-3P y Pyruvate en el mapa eco01100<sup>53</sup>**

<sup>53</sup> Figura 49. Fuente: [http://www.kegg.jp/kegg-bin/show\\_pathway?eco00900](http://www.kegg.jp/kegg-bin/show_pathway?eco00900)

Al unir los mapas entre sí, estos no sólo aportan nodos de distintos tipo, sino que también arcos que establecen una relación que permite la creación de puentes y así, nuevas rutas metabólicas.

*c. El registro de partes biológicas no se encuentra estandarizado*

El repositorio de partes biológicas de iGem contiene muestras de ADN de miles de partes que han sido agregadas por equipos y laboratorios participantes en la competición de iGem. Sin embargo, muchas de las piezas que se han subido al sitio web tienen información incompleta, es decir, no se encuentran debidamente documentadas. Esto representa una limitante para el proyecto, ya que impide la creación de filtros por algún criterio que permita la clasificación de las piezas para un fin en específico. La Figura 50 muestra un ejemplo de biobricks que no se encuentran debidamente documentados.

-?-	Name	Protein	Description	Direction	Uniprot	KEGG	E.C.	Substrate	Product	Length
1 ★	BBa_K118000		dxs coding sequence encoding 1-deoxyxylulose-5-phosphate synthase							1866
1 ★	BBa_K115050		A-coA -> AA-coA							1188

**Figura 50. Biobricks no documentados<sup>54</sup>**

*d. Mapas de KEGG manualmente anotados*

<sup>54</sup> Figura 50. Fuente: <http://parts.igem.org/Biosynthesis>

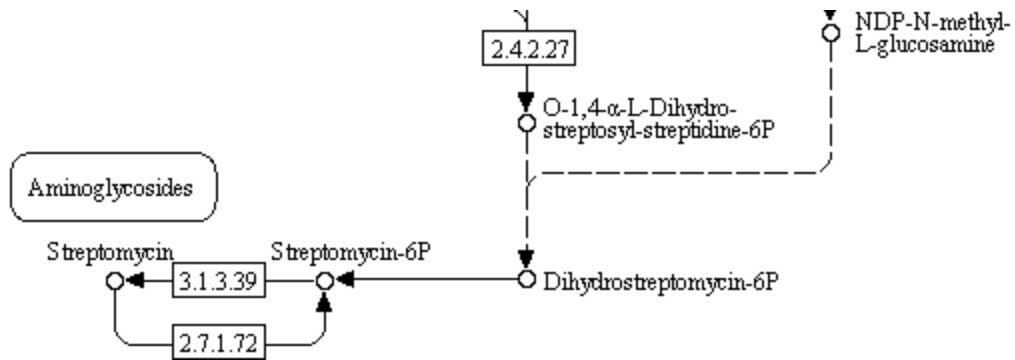
Los mapas de KEGG han sido manualmente anotados, por lo que la completitud de su información no se encuentra estandarizada. Los mapas se pueden procesar gracias al formato disponible KGML, sin embargo, al realizar la traducción de los mapas es necesario realizar consultas externas a los distintos repositorios de datos para completar los atributos que no se encuentra presentes. Un ejemplo de esto son las reacciones, en donde el archivo KGML no posee información de su nombre, sino que sólo el identificador asignado. Esto representa un obstáculo para el usuario, ya que debe consultar cada una de las reacciones para obtener una referencia más descriptiva.

Como una solución, se realizan consultas adicionales, ya sea en KEGG o en otras bases de datos, para obtener los datos adicionales requeridos.

*e. La ruta es dependiente del organismo que se seleccione*

Dado a que no todas las enzimas están presentes en todos los organismos, el grafo no contendrá todos los caminos enunciados en la ruta metabólica de referencia. Es decir, las conexiones existirán solamente entre aquellos nodos que se encuentran presentes. En el Apéndice 2 se encuentra el mapa para la biosíntesis de la estreptomicina, en el cual se muestra en color los nodos activos, por lo

tanto, la parte inferior, por ejemplo, no aparecerán como posibles caminos en el grafo.



**Figura 51. Elementos de las rutas metabólicas excluidas del grafo<sup>55</sup>**

f. *Delimitación de compuestos de entrada y salida*

Con el fin de acotar el alcance del proyecto, se definió un grupo limitado de entradas para llevar a cabo la mayoría de los análisis, como se menciona en la sección 3.3. De manera específica, estos fueron realizados principalmente con la Lactosa y el Piruvato, sin embargo, es posible variar los compuestos de entrada y salida, ya que al mismo tiempo, estas son entradas de la aplicación.

<sup>55</sup> Figura 51. Fuente: [http://www.genome.jp/kegg-bin/show\\_pathway?map00521](http://www.genome.jp/kegg-bin/show_pathway?map00521)

## **CAPITULO 6. CONCLUSIONES Y RECOMENDACIONES**

A pesar de que la Biología Sintética es una ciencia que se encuentra aún en desarrollo, desde sus inicios ha presentado resultados muy interesantes no sólo para la comunidad científica sino también para la sociedad.

La colaboración de los distintos grupos alrededor del mundo es esencial para el intercambio de ideas y el avance de la ciencia como tal en distintos escenarios. Es importante mencionar que estos avances no deben limitarse a un país ni un grupo de investigación en específico, sino ampliarse a acciones cooperativas multidisciplinarias.

Tal es el caso del prototipo desarrollado, que a pesar de encontrarse en un área en vías de desarrollo, representa un vital aporte para quienes si trabajan en Biología Sintética y desean llegar más allá de los logros alcanzados.

### **6.1. Conclusiones**

A continuación se enlistan puntualmente las conclusiones del proyecto:

1. El análisis de las rutas metabólicas de la bacteria E. Coli permitió el diseño de una herramienta de dominio específico para asistir el proceso de la producción de biocombustibles. A través de la



- traducción de estas rutas a grafos, utilizando la herramienta KEGGtranslator, y la aplicación de los algoritmos de búsqueda (Dijkstra y Búsqueda por anchura), se logró la creación y optimización de rutas entre la Lactosa y el Piruvato, a partir de una profundidad de 12 elementos en adelante. Esto habilitará el análisis de nuevos caminos para la generación sintética de biocombustibles.
2. El diseño de la herramienta de dominio específico se validó a través de la implementación de un prototipo funcional. Esto permitió la transformación y el análisis de más de 100 rutas metabólicas de la bacteria E. Coli y de la integración de los resultados con los repositorios de datos KEGG, Uniprot y GenBank. Con esto, se creó un catálogo de biobricks y se caracterizaron los ensamblajes finales, lo que permitirá el ensamblaje sintético de nuevos organismos para la producción de biocombustibles.
  3. Las caracterizaciones estructurales de los ensamblajes sintéticos finales se obtuvieron mediante la cuantificación del número de nucleótidos y aminoácidos de los ensamblajes, lo que enriquecerán la etapa de diseño de nuevas construcciones al proveer más información al usuario sobre la conformación del resultado final.

4. Las caracterizaciones funcionales de los ensamblajes se realizaron a través de la complementación de los resultados con el recurso universal de proteínas (UniProt), proporcionando información adicional acerca de las proteínas involucradas, tales como una descripción básica de sus funciones y las reacciones que catalizan, por ejemplo, "ATP + D-glucose = ADP + D-glucose 6-phosphate" para la proteína glucoquinasa, lo que permitirá orientar al usuario en la validación de las nuevas construcciones biológicas.
5. Las técnicas de representación gráfica utilizadas en el prototipo, permitieron la visualización de rutas metabólicas en grafos, a través del marco de trabajo JUNG, y los ensamblajes finales en gráficos circulares, a través de programación Web basada en JavaScript. Esto permitió mostrar los resultados de una manera sencilla, lo que facilitará su comprensión por parte del usuario.
6. La transformación de las rutas metabólicas en grafos hizo posible el análisis de datos por medio de la aplicación de teorías computacionales existentes, tales como estructuras de datos y algoritmos de búsqueda. Esto demuestra la aplicabilidad de la computación a un dominio no tradicional como la biología, así como la contribución de otros campos como la biología sintética y la computación de alto rendimiento.

7. La solución propuesta permite abstraer el proceso de rediseño de organismos al sugerir posibles rutas optimizadas entre los compuestos presentes en desechos orgánicos y en los biocombustibles, lo que permite al usuario enfocarse en problemas de mayor impacto y complejidad.
8. La colaboración entre profesionales del campo de la biotecnología y la computación fue esencial para el desarrollo del prototipo, tanto por la riqueza de sus ideas como por sus contribuciones desde el punto de vista de sus respectivas áreas de estudio.

A manera de resumen, finalmente se concluye que fue posible el diseño de una herramienta prototipo, la cual asistirá a los profesionales del área de la Biotecnología. Este prototipo permitirá el diseño de nuevas funcionalidades y el ensamblaje sintético de la bacteria E.Coli para la producción de biocombustibles, gracias a la implementación de las rutas metabólicas en estructuras computacionales y la validación por medio de criterio experto.

## 6.2. Recomendaciones

Además de las características desarrolladas existen algunas recomendaciones importantes que representan el trabajo futuro de esta investigación. A continuación, se detallan las posibles mejoras al prototipo actual:

- *Desarrollo de un traductor universal:* La información en las bases de datos genómicas se encuentran en formatos diferentes, lo que dificulta el análisis complementario de la información y la estandarización de los datos. Para este prototipo, se utilizaron las rutas almacenadas en KEGG y transformadas por medio del KEGGTranslator. Sin embargo, no sólo se presentan las limitaciones mencionadas en la sección 5.4.1, sino que también excluye la posibilidad de incluir mapas de otras bases de datos como Metacyc (Caspi, Altman, & Billington, 2014) y Reactome (Croft, O'Kelly, & Wu, 2011), que utilizan el formato SBML. Como alternativa, se propone incluir distintos traductores para que sea compatible con el conjunto de formatos más relevantes en este campo.
- *Creación de una interfaz web:* Para la implementación de la propuesta presentada se tomó la decisión de desarrollar el flujo por medio de la aplicación en línea de comandos, la cual

funcionalmente facilita la validación de los resultados. Sin embargo, desde un punto de vista de usabilidad para el usuario final, es posible abstraer aún más las funciones y requerimientos en cada paso, razón por la cual, se propone el desarrollo de una interfaz web, de modo que pueda ser accesible a otros usuarios potenciales por medio de las infraestructuras disponibles, como el Clúster Nelly del Proyecto de Bioinformática de la Escuela de Medicina de la Universidad de Costa Rica (Universidad de Costa Rica, 2011).

- *Participación de usuario:* debido a que existe un grupo limitado de profesionales trabajando en el país para esta área, la etapa de validación fue difícil debido a los ciclos de retroalimentación y el tiempo disponible por parte de los usuarios. Sin embargo, fue esencial la participación de los usuarios finales para la mejora del sistema de acuerdo a sus necesidades específicas.
- *Utilización de varias bases genómicas:* a pesar de que KEGG provee suficiente información para establecer rutas de un compuesto a otro, sería relevante complementar la información de otros repositorios de datos, tanto como para validar la relevancia de las rutas así también para completar datos que se encuentren incompletos. También, es posible llevar a cabo análisis estadísticos

sobre la aplicación de la metodología diseñada, utilizando las rutas metabólicas almacenadas en las diferentes bases de datos.

- *Escalabilidad de la aplicación:* el prototipo diseñado permitió obtener una cantidad significativa de resultados acerca de las posibles rutas entre la Lactosa y el Piruvato, para el aprovechamiento de desechos orgánicos y la producción de biocombustibles. Sin embargo, en otros escenarios en donde se requieran análisis a mayor escala, es decir, a más de 16 niveles de profundidad y grafos de más de 4.000 nodos, se deberá realizar una reevaluación de los algoritmos y proponer diferentes enfoques que permitan así extender el alcance de la aplicación.

Finalmente, para dar a conocer y compartir los resultados de la presente investigación, así como para buscar usuarios potenciales del prototipo, se desea realizar una publicación en alguna revista científica relacionada como parte del trabajo futuro de este proyecto.

## BIBLIOGRAFÍA

- Aguilar, J. G. (15 de Octubre de 2007). *Ingeniería en Biotecnología celebra diez años de logros*. Recuperado el 01 de Mayo de 2013, de <http://www.tec.cr/prensa/informatec/2007/OctubreI/n17.htm>
- Alvarado Meza, R. (03 de Marzo de 2014). Proceso de selección de proteínas. (L. Vásquez, Entrevistador) San José.
- Alvarado, R., García, D., & Rodríguez, A. (2012). *Team: Costa Rica-TEC-UNA Project iGEM*. Recuperado el 2013 de Junio de 23, de [http://2012.igem.org/Team:Costa\\_Rica-TEC-UNA/Project](http://2012.igem.org/Team:Costa_Rica-TEC-UNA/Project)
- Atsumi, S., Hanai, T., & Liao, J. C. (3 de Enero de 2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature*, *451*, 86-89.
- Benson, D. A., Cavanaugh, M., & Clark, K. (26 de Noviembre de 2012). GenBank. *Nucleic Acids Research*, D36–D42. Recuperado el 16 de 03 de 2014, de <http://www.ncbi.nlm.nih.gov/genbank/>
- Bolchini, D., Finkelstein, A., Perrone, V., & Nagl, S. (9 de Diciembre de 2009). Better bioinformatics through usability analysis. *Bioinformatics*, *25*(3), 406-412.
- Caspi, R., Altman, T., & Billington, R. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, *42*(D1), D459–D471. Recuperado el 5 de Mayo de 2014, de <http://www.metacyc.org/>
- Chin, M. (2 de Enero de 2008). *UCLA researchers develop method for production of more efficient biofuels*. (UCLA Newsroom) Recuperado el 7 de Julio de 2013, de <http://newsroom.ucla.edu/portal/ucla/ucla-engineering-researchers-develop-42502.aspx>
- Consejo Nacional para Investigaciones Científicas y Tecnológicas. (29 de Enero de 2004). *UNU-BIOLAC\_CONICIT*. Recuperado el 23 de Agosto de 2013, de [http://www.conicit.go.cr/sector\\_cyt/convenios\\_cooperacion/UNU-BIOLAC\\_CONICIT.html](http://www.conicit.go.cr/sector_cyt/convenios_cooperacion/UNU-BIOLAC_CONICIT.html)
- Croft, D., O'Kelly, G., & Wu, G. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids*

- Research*, 30, D691–D697. Recuperado el 5 de Mayo de 2014, de <http://www.reactome.org/>
- Demirbas, A. (Agosto de 2008). Biofuels sources, biofuel policy, biofuel economy and global biofuel projections. *Energy Conversion and Management*, 49(8), 2106-2116.
- Feltman, R. (22 de Abril de 2013). Hijacking E. Coli to Brew Synthetic Fuel. *Popular Mechanics*.
- Fielding, R. T. (2000). *Fielding Dissertation: CHAPTER 5: Representational State Transfer (REST)*. Recuperado el 8 de Marzo de 2014, de Architectural Styles and the Design of Network-based Software Architectures: [https://www.ics.uci.edu/~fielding/pubs/dissertation/rest\\_arch\\_style.htm](https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm)
- Fuga, J., & Collier, V. (2013). Thirteenth Annual Freshman Engineering Conference. *Genetic And Metabolic Engineering Of Escherichia Coli For Biofuel Production As An Alternative Fuel Source*. Pittsburgh, PA.
- Google Developers. (21 de Enero de 2014). *Visualization: Pie Charts*. (Google) Recuperado el 2014 de Mayo de 17, de Google Charts: <https://developers.google.com/chart/interactive/docs/gallery/pie-chart>
- GraphML Project Group. (2007). *The GraphML File Format*. Recuperado el 08 de Marzo de 2014, de <http://graphml.graphdrawing.org/>
- Hayden, E. C. (31 de Agosto de 2013). *Popular plant database set to charge users: Nature News & Comment*. Recuperado el 1 de Mayo de 2013, de <http://www.nature.com/news/popular-plant-database-set-to-charge-users-1.13642>
- Heinemann, M., & Panke, S. (29 de Agosto de 2006). Synthetic biology— Putting engineering into biology. *Bioinformatics*, 22(22), 2790-2799.
- Hill, A. D., Tomshine, J. R., Weeding, E. M., Sotiropoulos, V., & Kaznessis, Y. N. (30 de Agosto de 2008). SynBioSS: the synthetic biology modeling suit. *Bioinformatics*, 24(21), 2551-2553.
- Howard, T. P., Middelhaufe, S., & Moore, K. (23 de Abril de 2013). Synthesis of customized petroleum-replica fuel molecules by targeted modification of free fatty acid pools in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United*



- States of America (PNAS)*, 110(19), 7636-7641. Recuperado el 7 de Julio de 2013, de <http://www.greencarcongress.com/2013/04/exeter-20130423.html>
- Hu, J. (14 de Junio de 2013). *Escherichia coli - Ecoliwiki*. Recuperado el 23 de Junio de 2013, de [http://ecoliwiki.net/colipedia/index.php?title=Escherichia\\_coli&oldid=1400336](http://ecoliwiki.net/colipedia/index.php?title=Escherichia_coli&oldid=1400336)
- iGem Foundation. (26 de Junio de 2014). *Biosynthesis - parts.igem.org*. Obtenido de <http://parts.igem.org/Biosynthesis>
- INBio. (2011). *Atta 2.0*. Recuperado el 3 de Julio de 2013, de <http://pulsatrix.inbio.ac.cr/projects/atta2/>
- International Genetically Engineered Machine Foundation. (2003). *Assembly:Standard assembly - parts.igem.org*. Recuperado el 3 de Mayo de 2014, de Registry of Standard Biological Parts: [http://parts.igem.org/Assembly:Standard\\_assembly](http://parts.igem.org/Assembly:Standard_assembly)
- International Genetically Engineered Machine Foundation. (2003). *Help:Prefix-Suffix - parts.igem.org*. Recuperado el 3 de Mayo de 2014, de Registry of Standard Biological Parts: <http://parts.igem.org/Help:Prefix-Suffix>
- International Genetically Engineered Machine Foundation. (2003). *Help:Restriction enzymes - parts.igem.org*. Recuperado el 27 de Abril de 2014, de Registry of Standard Biological Parts: [http://parts.igem.org/Help:Restriction\\_enzymes](http://parts.igem.org/Help:Restriction_enzymes)
- International Genetically Engineered Machine Foundation. (2003). *Main Page - parts.igem.org*. Recuperado el 2014 de Mayo de 03, de Registry of Standard Biological Parts: [http://parts.igem.org/Main\\_Page](http://parts.igem.org/Main_Page)
- Kanehisa Laboratories. (2014). *KEGG API*. Recuperado el 8 de Marzo de 2014, de <http://www.kegg.jp/kegg/docs/keggapi.html>
- Kanehisa Laboratories. (28 de Abril de 2014). *KEGG PATHWAY Database*. Recuperado el 04 de 04 de 2014, de <http://www.genome.jp/kegg/pathway.html>
- Kanehisa Laboratories. (2014). *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Recuperado el 8 de Marzo de 2014, de <http://www.kegg.jp/>

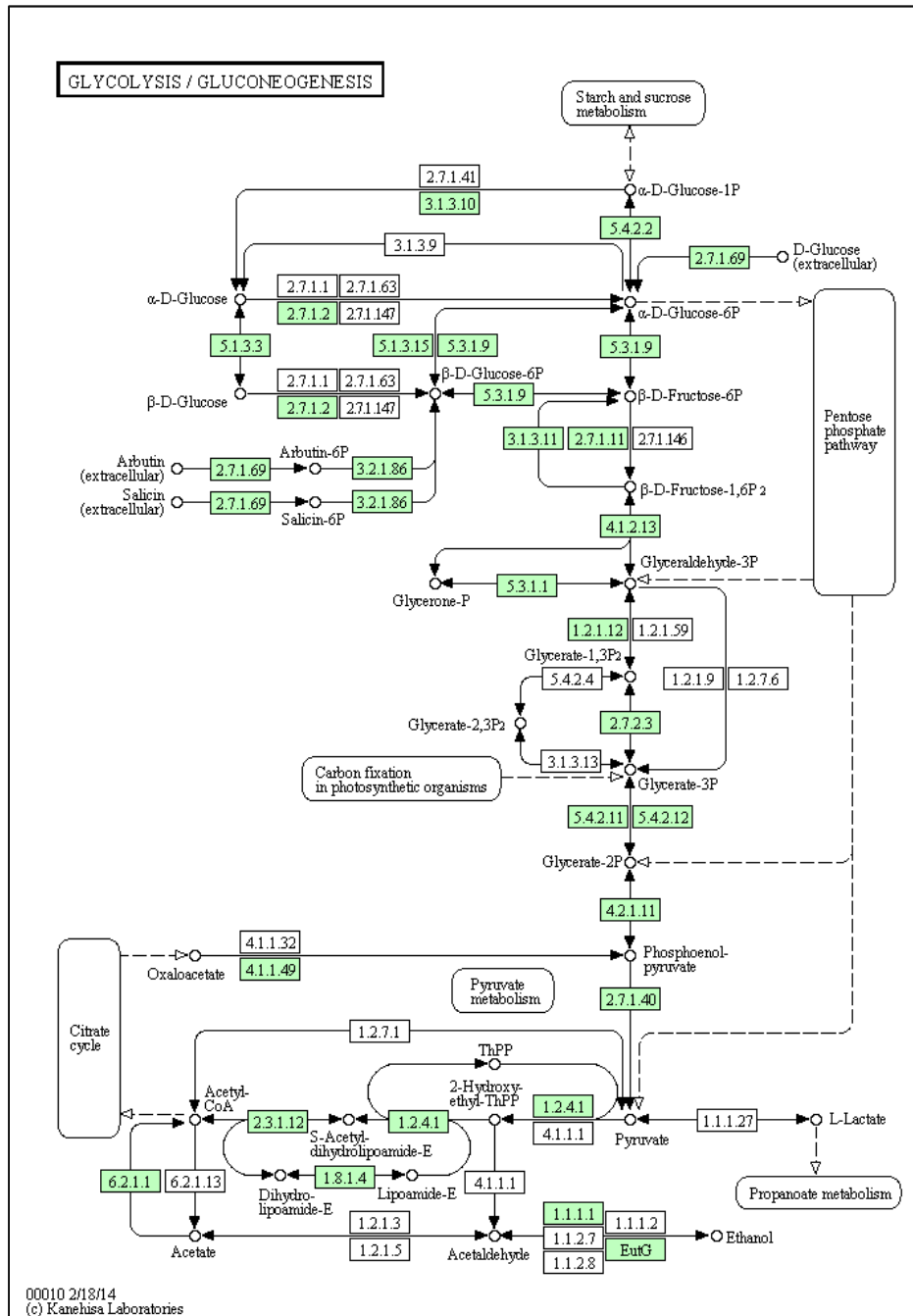
- Kanehisa Laboratories. (2014). *KGML (KEGG Markup Language)*. Recuperado el 8 de Marzo de 2014, de <http://www.kegg.jp/kegg/xml/>
- Kollewe, J. (17 de Febrero de 2012). DNA machine can sequence human genomes in hours. *The Guardian*. Recuperado el 7 de Julio de 2013, de <http://www.guardian.co.uk/science/2012/feb/17/dna-machine-human-sequencing>
- Liu, T., & Khosla, C. (2010). Genetic Engineering of Escherichia coli for Biofuel Production. *Annual Review of Genetics*, 44, 53-69.
- Lorimer, D., Raymond, A., & Walchli, J. (21 de Abril de 2009). Gene composer: database software for protein construct design, codon engineering, and gene synthesis. *BMC Biotechnology*, 9, 36.
- Matarrita Araya, F. (2013). *Alineamiento de Secuencias*. Obtenido de <http://bioinformatica.ucr.ac.cr/needleman/>
- Morales, A., & Gurza, F. (Diciembre de 2002). *Elaboración de una Bebida como Alternativa al Manejo de Suero Lácteo*. Obtenido de Universidad Earth: <http://usi.earth.ac.cr/glas/sp/pdf/99067.pdf>
- Morales, J. C. (29 de Julio de 2013). Aplicaciones de Bioinformática en el país. *Entrevista CeNat*. San José.
- Morris, J. (1998). *Data Structure and Algorithms: Dijkstra's Algorithm*. Recuperado el 14 de 03 de 2014, de <https://www.cs.auckland.ac.nz/software/AlgAnim/dijkstra.html>
- National Center for Biotechnology Information. (1994). *E. coli chromosomal region from 76.0 to 81.5 minutes - Nucleotide - NCBI*. (U.S. National Library of Medicine) Recuperado el 17 de Mayo de 2014, de <http://www.ncbi.nlm.nih.gov/nucore/U00039.1>
- National Center for Biotechnology Information. (2010). *Entrez Programming Utilities Help - NCBI Bookshelf*. Recuperado el 16 de 03 de 2014, de <http://www.ncbi.nlm.nih.gov/books/NBK25501/>
- National Center for Biotechnology Information. (09 de Agosto de 2013). *E-utilities Quick Start - Entrez Programming Utilities Help - NCBI Bookshelf*. Obtenido de [http://www.ncbi.nlm.nih.gov/books/NBK25500/#chapter1.Demonstration\\_Programs](http://www.ncbi.nlm.nih.gov/books/NBK25500/#chapter1.Demonstration_Programs)

- Nature Education. (2013). *plasmid / plasmids*. Obtenido de Learn Science at Scitable: <http://www.nature.com/scitable/definition/plasmid-plasmids-28>
- Orozco Solano, A. (2012). Bioinformática en Costa Rica. En *Informe 2012: Hacia la Sociedad de la Información* (págs. 229-256). San José: PROSIC, Universidad de Costa Rica.
- Orozco, A., Morera, J., Jiménez, S., & Boza, R. (30 de Mayo de 2013). A review of Bioinformatics training applied to research in Molecular Medicine, Agriculture and Biodiversity in Costa Rica and Central America. *Briefings in Bioinformatics*, 14(5), 661-700.
- Pavelin, K., Cham, J. A., & Matos, P. d. (26 de Junio de 2014). *Bioinformatics Meets User-Centred Design: A Perspective*. Obtenido de PLOS Computational Biology: <http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002554>
- Perkel, J. (Febrero de 2013). Finding the true \$1000 genome. *BioTechniques: The International Journal of Life Science Methods*, 54(2), 71-74.
- Piacente, P. J. (6 de Octubre de 2011). *La biología sintética impulsa el desarrollo de los biocombustibles*. Recuperado el 23 de Junio de 2013, de [http://www.tendencias21.net/La-biologia-sintetica-impulsa-el-desarrollo-de-los-biocombustibles\\_a7761.html](http://www.tendencias21.net/La-biologia-sintetica-impulsa-el-desarrollo-de-los-biocombustibles_a7761.html)
- Rahman, S., Advani, P., & Schunk, R. (2005). Metabolic pathway analysis web service - Pathway Hunter Tool at CUBIC. 7(21).
- Rodriguez, M. A. (13 de Junio de 2012). *Property Graph Model - tinkerpops/blueprints - GitHub*. Recuperado el 8 de Marzo de 2014, de <https://github.com/tinkerpops/blueprints/wiki/Property-Graph-Model>
- Rosales, C. (20 de Junio de 2013). Biocombustibles en el CENIBiot. (L. Vasquez, Entrevistador)
- Salas, D. L. (27 de Junio de 2013). Bioinformática permitirá desarrollo de tratamientos y fármacos personalizados en Costa Rica. *El Financiero*.
- Sayers, E. (2010). *Ebot*. Recuperado el 16 de 03 de 2014, de <http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi>

- Synthetic Biology Community. (2013). *Synthetic Biology: FAQ*. Recuperado el 2013 de Junio de 23, de <http://syntheticbiology.org/FAQ.html>
- The Institute of Grocery Distribution. (04 de Junio de 2007). *Biofuels*. Recuperado el 23 de Junio de 2013, de <http://www.igd.com/our-expertise/Sustainability/Energy/3435/Biofuels/>
- The JUNG Framework Development Team. (2010). *JUNG - Java Universal Network/Graph Framework*. Recuperado el 14 de 03 de 2014, de <http://jung.sourceforge.net/>
- TinkerPop team. (2009). *Open Source Software Products in the Graph Space*. Recuperado el 8 de Marzo de 2014, de TinkerPop: <http://www.tinkerpop.com/>
- UniProt Consortium. (28 de Mayo de 2012). *How can I access resources on this website programmatically?* Recuperado el 19 de Mayo de 2014, de UniProt FAQ: [http://www.uniprot.org/faq/28#id\\_mapping\\_java\\_example](http://www.uniprot.org/faq/28#id_mapping_java_example)
- UniProt Consortium. (2014). *UniProt*. Recuperado el 16 de 03 de 2014, de <http://www.uniprot.org/>
- Universidad de Costa Rica. (2011). *Galaxy - Cluster Nelly*. Recuperado el 1 de Mayo de 2014, de <http://www.bioinformatica.ucr.ac.cr:8080/>
- Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J., & Govindarajan, S. (6 de Junio de 2006). Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics*, 7, 285.
- Wrzodek, C., Dräger, A., & Zell, A. (2011). KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics*, 27(16), 2314-2315. Recuperado el 8 de Marzo de 2014, de <http://www.ra.cs.uni-tuebingen.de/software/KEGGtranslator/index.htm>

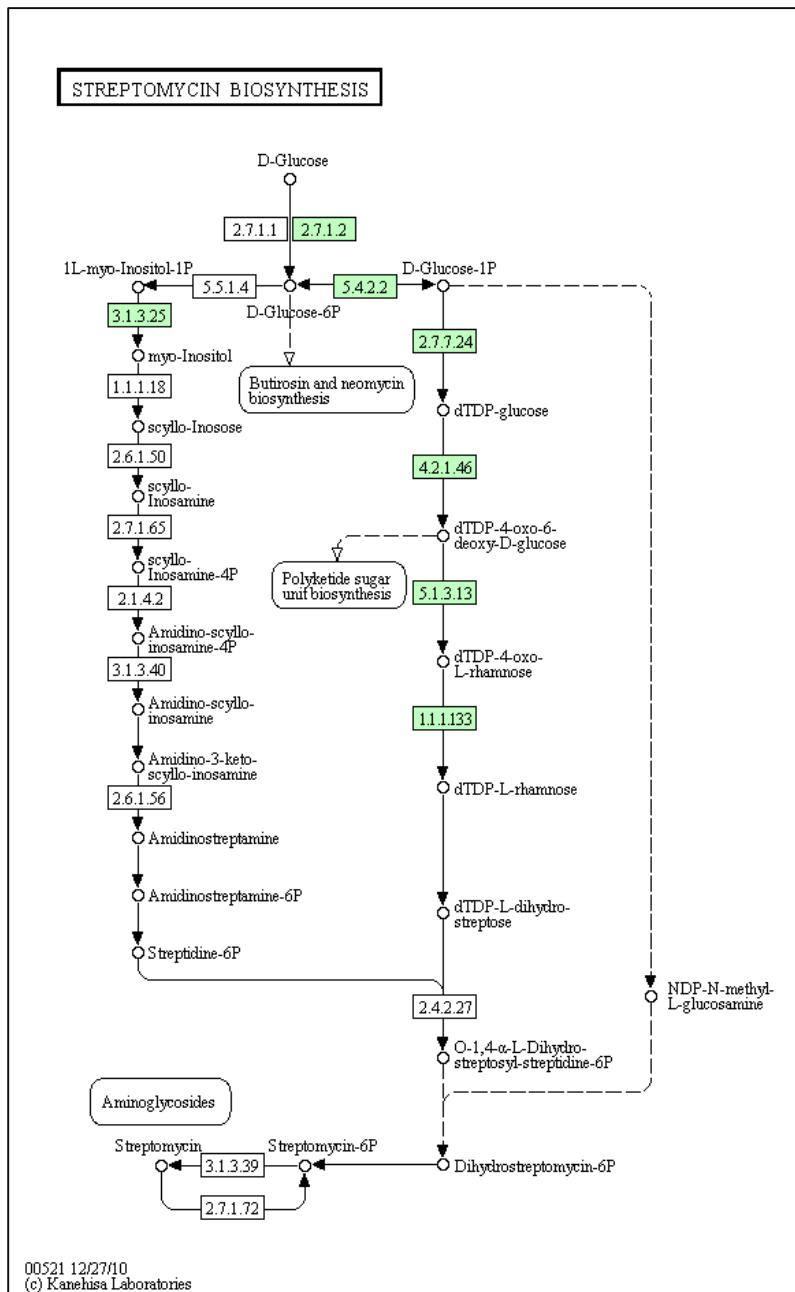
# APÉNDICES

## Apéndice 1. Mapa de la glucólisis en la bacteria *E. Coli*<sup>56</sup>



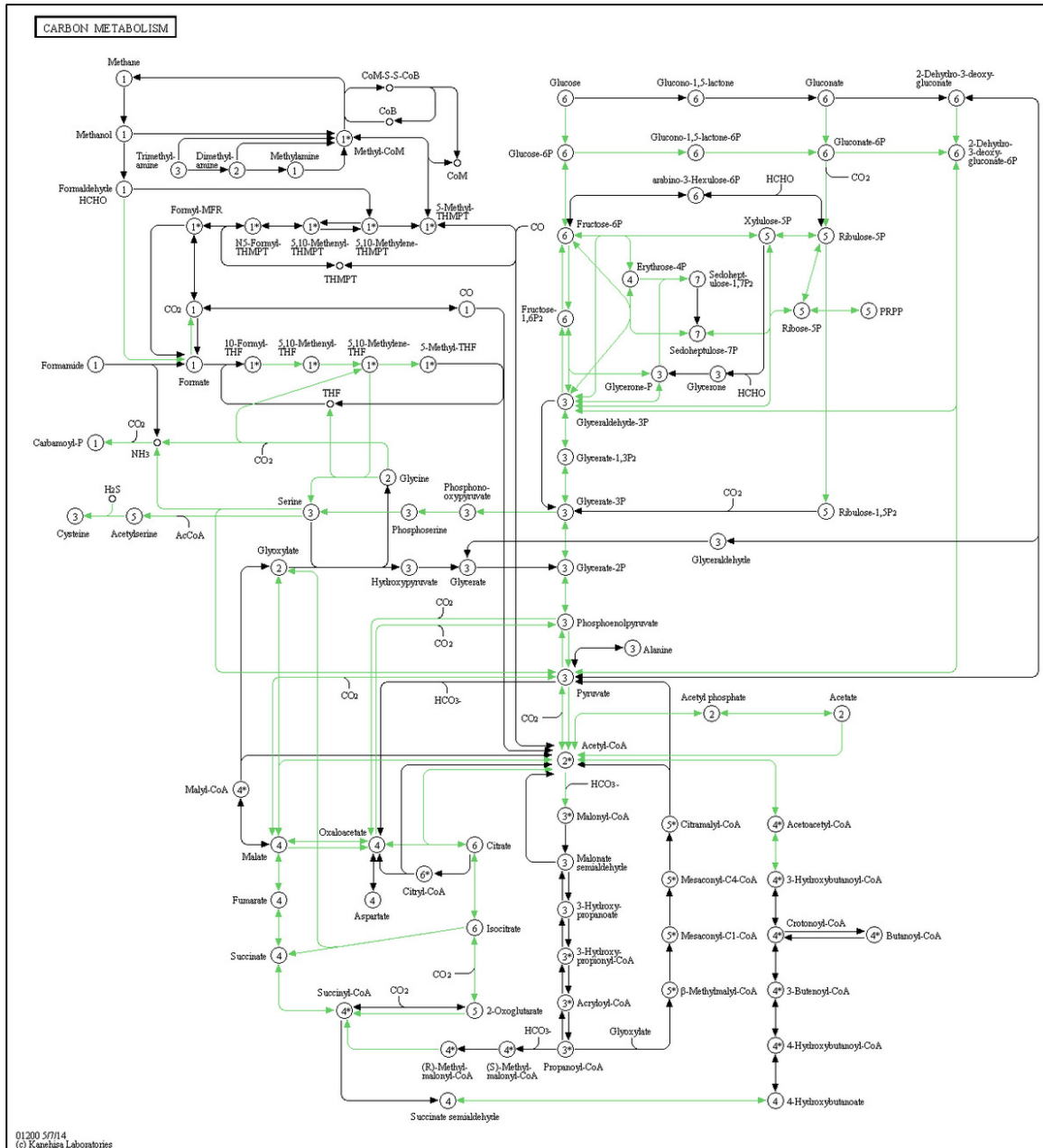
<sup>56</sup> Apéndice 1. Fuente: [http://www.genome.jp/kegg-bin/show\\_pathway?eco00010](http://www.genome.jp/kegg-bin/show_pathway?eco00010)

## Apéndice 2. Mapa de la biosíntesis de la estreptomicina en la bacteria *E. Coli*<sup>57</sup>



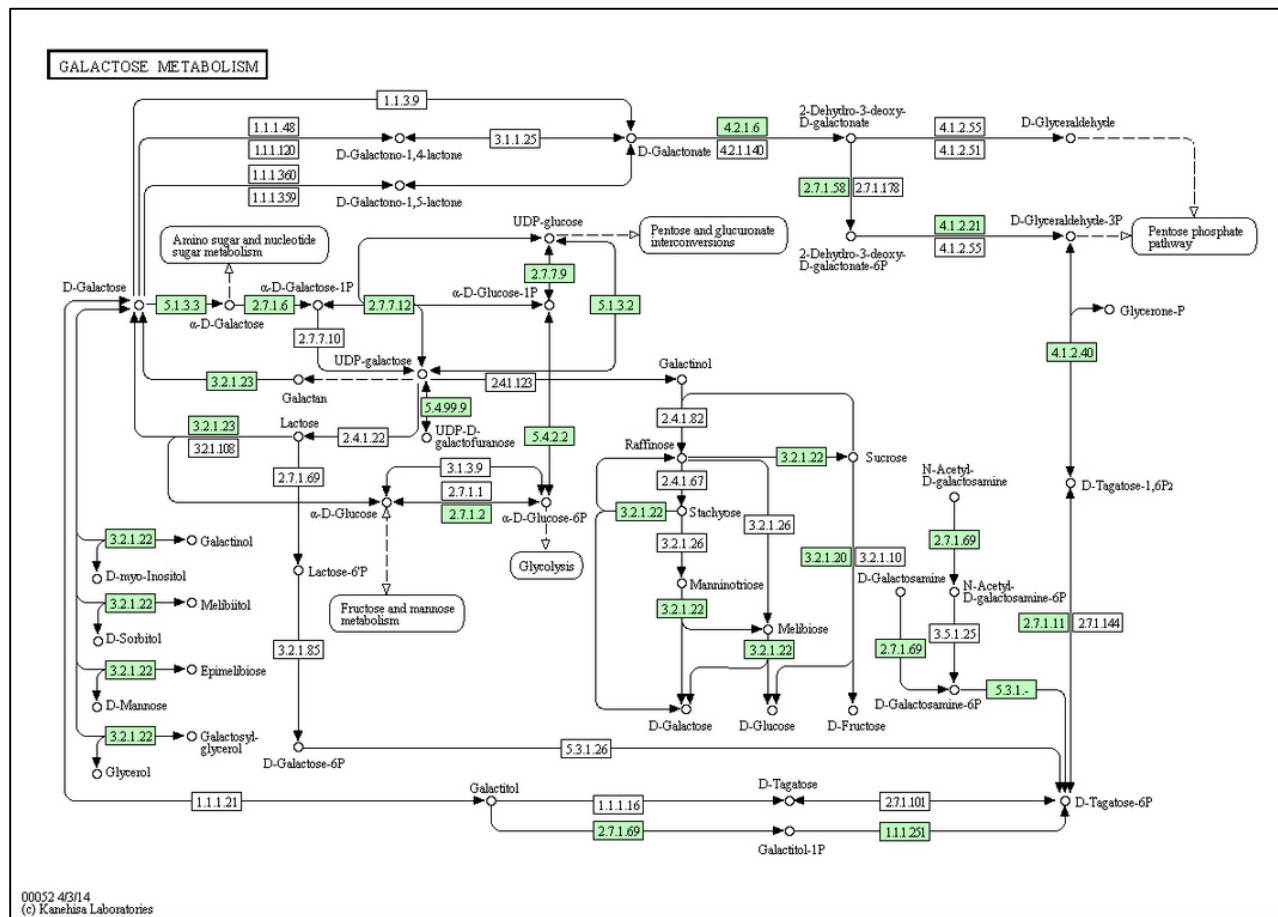
<sup>57</sup> Apéndice 2. Fuente: [http://www.genome.jp/kegg-bin/show\\_pathway?eco00521](http://www.genome.jp/kegg-bin/show_pathway?eco00521)

# Apéndice 3. Mapa del metabolismo del carbono en la bacteria *E. Coli*<sup>58</sup>



<sup>58</sup> Apéndice 3. Fuente: [http://www.genome.jp/kegg-bin/show\\_pathway?eco01200](http://www.genome.jp/kegg-bin/show_pathway?eco01200)

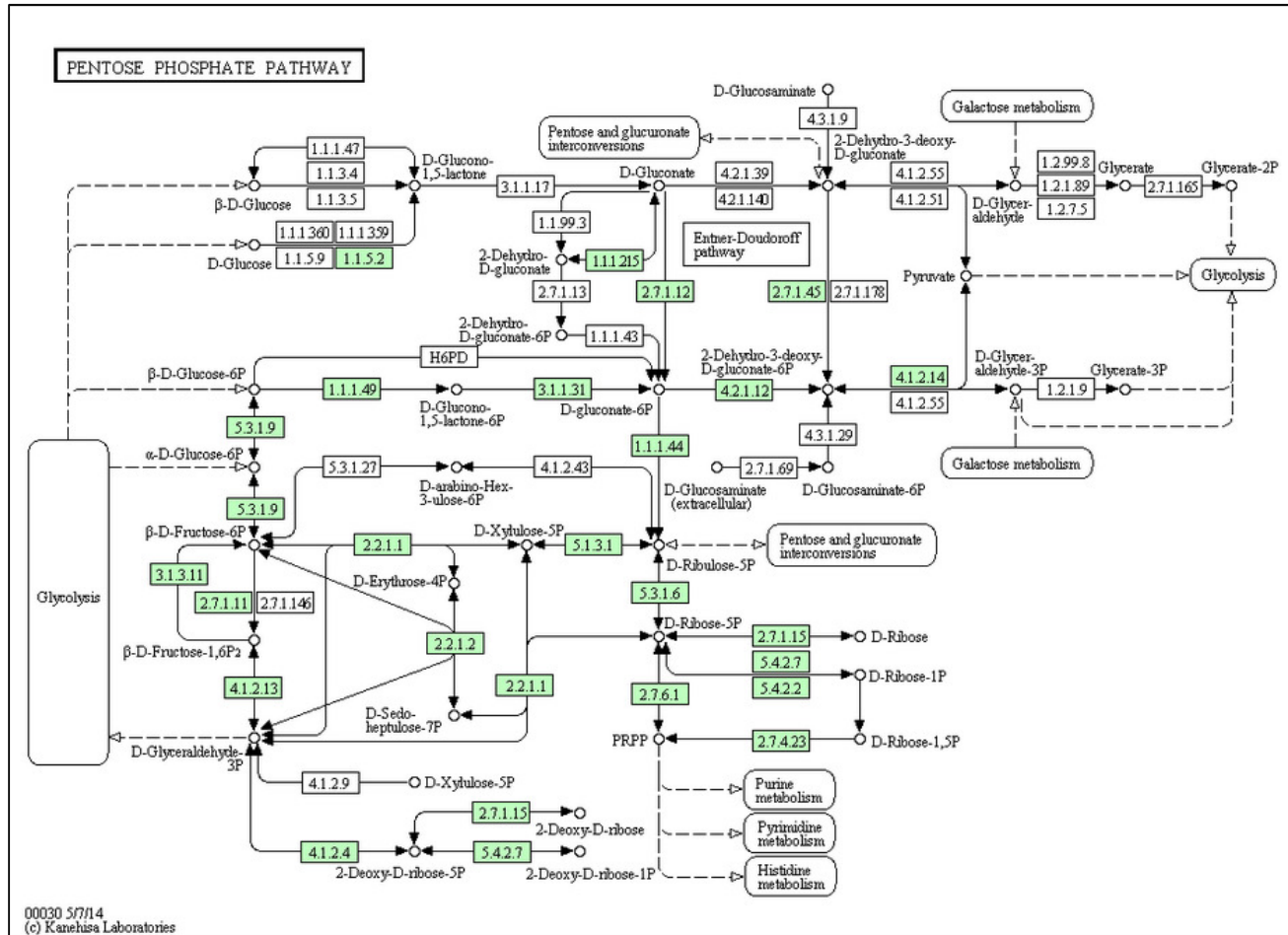
## Apéndice 4. Mapa del metabolismo de la galactosa en la bacteria *E. Coli*<sup>59</sup>



<sup>59</sup> Apéndice 4. Fuente: [http://www.genome.jp/kegg-bin/show\\_pathway?eco00052](http://www.genome.jp/kegg-bin/show_pathway?eco00052)



## Apéndice 5. Mapa de la ruta de la pentosa fosfato en la bacteria *E. Coli*<sup>60</sup>



<sup>60</sup> Apéndice 5. Fuente: [http://www.genome.jp/kegg-bin/show\\_pathway?eco00030](http://www.genome.jp/kegg-bin/show_pathway?eco00030)

## Apéndice 6. Archivo en formato KGML de la glucólisis en bacteria E. Coli<sup>61</sup>

```
▼<!--
  Creation date: Feb 18, 2014 11:53:54 +0900 (GMT+09:00)
-->
▼<pathway name="path:eco00010" org="eco" number="00010" title="Glycolysis / Gluconeogenesis"
  image="http://www.kegg.jp/kegg/pathway/eco/eco00010.png" link="http://www.kegg.jp/kegg-bin/show_pathway?eco00010">
  ▼<entry id="13" name="eco:b2097 eco:b2925" type="gene" reaction="rn:R01070" link="http://www.kegg.jp/dbget-bin/www_bget?
    eco:b2097+eco:b2925">
    <graphics name="fbaB..." fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle" x="483" y="404" width="46" height="17"/>
  </entry>
  ▼<entry id="37" name="ko:K00128 ko:K14085 ko:K00149" type="ortholog" reaction="rn:R00710" link="http://www.kegg.jp/dbget-
    bin/www_bget?K00128+K14085+K00149">
    <graphics name="K00128..." fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="289" y="943" width="46" height="17"/>
  </entry>
  ▼<entry id="38" name="ko:K01905" type="ortholog" reaction="rn:R00229" link="http://www.kegg.jp/dbget-bin/www_bget?K01905">
    <graphics name="K01905" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="146" y="911" width="46" height="17"/>
  </entry>
  ▼<relation entry1="70" entry2="73" type="ECrel">
    <subtype name="compound" value="87"/>
  </relation>
  ▼<relation entry1="69" entry2="70" type="ECrel">
    <subtype name="compound" value="87"/>
  </relation>
  ▼<relation entry1="69" entry2="73" type="ECrel">
    <subtype name="compound" value="87"/>
  </relation>
</pathway>
```

---

<sup>61</sup> Apéndice 6. Fuente: <http://www.kegg.jp/kegg-bin/download?entry=eco00010&format=kgml>

## Apéndice 7. Descarga de archivos KGML<sup>62</sup>

```
#!/usr/bin/ruby

class DownloadKegg

  def initialize args

    @file = args[0]
    @prefix = "eco"

  end

  def readMaps

    @maps = File.readlines(@file);

    @maps.map! { |item|
      item.strip!
      item.gsub! ("map", @prefix)
      puts "wget http://rest.kegg.jp/get/#{item}/kgml -O #{item}.xml"
      %x(wget http://rest.kegg.jp/get/#{item}/kgml -O #{item}.xml)
    }

  end

  def start
    readMaps
  end

end

if __FILE__ == $0
  downloader = DownloadKegg.new(ARGV)
  downloader.start
end
```

---

<sup>62</sup> Apéndice 7. Elaboración propia.

## Apéndice 8. Lista de rutas metabólicas para la Lactosa y Piruvato.<sup>63</sup>

- map00010 Glycolysis / Gluconeogenesis.
- map00020 Citrate cycle (TCA cycle).
- map00030 Pentose phosphate pathway.
- map00040 Pentose and glucuronate interconversions.
- map00051 Fructose and mannose metabolism.
- map00052 Galactose metabolism.
- map00053 Ascorbate and aldarate metabolism.
- map00061 Fatty acid biosynthesis.
- map00071 Fatty acid degradation.
- map00130 Ubiquinone and other terpenoid-quinone biosynthesis.
- map00190 Oxidative phosphorylation.
- map00240 Pyrimidine metabolism.
- map00250 Alanine, aspartate and glutamate metabolism.
- map00260 Glycine, serine and threonine metabolism.
- map00270 Cysteine and methionine metabolism.
- map00280 Valine, leucine and isoleucine degradation.
- map00281 Geraniol degradation.
- map00290 Valine, leucine and isoleucine biosynthesis.
- map00300 Lysine biosynthesis.
- map00310 Lysine degradation.
- map00330 Arginine and proline metabolism.
- map00340 Histidine metabolism.
- map00350 Tyrosine metabolism.
- map00361 Chlorocyclohexane and chlorobenzene degradation.
- map00362 Benzoate degradation.
- map00364 Fluorobenzoate degradation.
- map00380 Tryptophan metabolism.
- map00400 Phenylalanine, tyrosine and tryptophan biosynthesis.
- map00410 beta-Alanine metabolism.
- map00430 Taurine and hypotaurine metabolism.

---

<sup>63</sup> Apéndice 8. Fuente:

[http://www.kegg.jp/kegg-bin/search\\_pathway\\_text?map=map&keyword=pyruvate&mode=1](http://www.kegg.jp/kegg-bin/search_pathway_text?map=map&keyword=pyruvate&mode=1)  
[http://www.kegg.jp/kegg-bin/search\\_pathway\\_text?map=map&keyword=lactose&mode=1](http://www.kegg.jp/kegg-bin/search_pathway_text?map=map&keyword=lactose&mode=1)

- map00440 Phosphonate and phosphinate metabolism.
- map00450 Selenocompound metabolism.
- map00460 Cyanoamino acid metabolism.
- map00471 D-Glutamine and D-glutamate metabolism.
- map00473 D-Alanine metabolism.
- map00500 Starch and sucrose metabolism.
- map00511 Other glycan degradation.
- map00520 Amino sugar and nucleotide sugar metabolism.
- map00521 Streptomycin biosynthesis.
- map00523 Polyketide sugar unit biosynthesis.
- map00540 Lipopolysaccharide biosynthesis.
- map00550 Peptidoglycan biosynthesis.
- map00561 Glycerolipid metabolism.
- map00562 Inositol phosphate metabolism.
- map00564 Glycerophospholipid metabolism.
- map00565 Ether lipid metabolism.
- map00590 Arachidonic acid metabolism.
- map00592 alpha-Linolenic acid metabolism.
- map00600 Sphingolipid metabolism.
- map00620 Pyruvate metabolism.
- map00621 Dioxin degradation.
- map00622 Xylene degradation.
- map00623 Toluene degradation.
- map00625 Chloroalkane and chloroalkene degradation.
- map00626 Naphthalene degradation.
- map00627 Aminobenzoate degradation.
- map00633 Nitrotoluene degradation.
- map00640 Propanoate metabolism.
- map00642 Ethylbenzene degradation.
- map00650 Butanoate metabolism.
- map00660 C5-Branched dibasic acid metabolism.
- map00670 One carbon pool by folate.
- map00680 Methane metabolism.
- map00740 Riboflavin metabolism.
- map00750 Vitamin B6 metabolism.
- map00760 Nicotinate and nicotinamide metabolism.
- map00770 Pantothenate and CoA biosynthesis.
- map00785 Lipoic acid metabolism.

- map00790 Folate biosynthesis.
- map00860 Porphyrin and chlorophyll metabolism.
- map00900 Terpenoid backbone biosynthesis.
- map00903 Limonene and pinene degradation.
- map00910 Nitrogen metabolism.
- map00920 Sulfur metabolism.
- map00930 Caprolactam degradation.
- map00970 Aminoacyl-tRNA biosynthesis.
- map01040 Biosynthesis of unsaturated fatty acids.
- map01053 Biosynthesis of siderophore group nonribosomal peptides.
- map01100 Metabolic pathways.
- map01110 Biosynthesis of secondary metabolites.
- map01120 Microbial metabolism in diverse environments.
- map01200 Carbon metabolism.
- map01210 2-Oxocarboxylic acid metabolism.
- map01212 Fatty acid metabolism.
- map01220 Degradation of aromatic compounds.
- map01230 Biosynthesis of amino acids.
- map02010 ABC transporters.
- map02020 Two-component system.
- map02030 Bacterial chemotaxis.
- map02040 Flagellar assembly.
- map02060 Phosphotransferase system (PTS).
- map03010 Ribosome.
- map03018 RNA degradation.
- map03020 RNA polymerase.
- map03030 DNA replication.
- map03060 Protein export.
- map03070 Bacterial secretion system.
- map03410 Base excision repair.
- map03420 Nucleotide excision repair.
- map03430 Mismatch repair.
- map03440 Homologous recombination.
- map04122 Sulfur relay system.

## **Apéndice 9. Categorías y biobricks seleccionados (iGem Foundation, 2014)**

### **Degradation of AHL**

- BBa\_C0061
- BBa\_C0060
- BBa\_C0070
- BBa\_C0076
- BBa\_C0078
- BBa\_C0161
- BBa\_C0170
- BBa\_C0178
- BBa\_K091109
- BBa\_C0060
- BBa\_C0160

### **Isoprenoids**

- BBa\_K118000
- BBa\_K115050
- BBa\_K115056
- BBa\_K115057
- BBa\_K118002
- BBa\_K118003
- BBa\_K118008
- BBa\_K343001
- BBa\_K849000
- BBa\_K849001
- BBa\_K849003

### **Odorants**

- BBa\_J45001
- BBa\_J45002
- BBa\_J45004
- BBa\_J45008
- BBa\_J45014
- BBa\_J45017
- BBa\_I742107

### **Plastic**

- BBa\_K125504
- BBa\_K125501
- BBa\_K125502
- BBa\_K125503
- BBa\_K156012
- BBa\_K156013
- BBa\_K156014
- BBa\_K759004
- BBa\_K759005

### **Butanol**

- BBa\_I725011
- BBa\_I72512
- BBa\_I725013
- BBa\_I725014
- BBa\_I725015

### **Bisphenol-A**

- BBa\_K123001
- BBa\_K123000
- BBa\_K525007
- BBa\_K525515
- BBa\_K525008
- BBa\_K525517
- BBa\_K525582
- BBa\_K525562

### **Cellulose**

- BBa\_K118022
- BBa\_K118023

- BBa\_K118028

**Other Enzymes:**

- BBa\_C0083
- BBa\_I15008
- BBa\_I15009
- BBa\_T9150
- BBa\_I716153
- BBa\_I716154
- BBa\_I716155
- BBa\_I716152
- BBa\_I742141
- BBa\_I742142
- BBa\_I723024
- BBa\_I723025
- BBa\_K137005
- BBa\_K137006
- BBa\_K137009
- BBa\_K137011
- BBa\_K137017
- BBa\_K118015
- BBa\_K118016
- BBa\_K123001
- BBa\_K108018
- BBa\_K108026
- BBa\_K108027
- BBa\_K108028
- BBa\_K108029
- BBa\_K147003
- BBa\_K123000
- BBa\_K356000
- BBa\_K417000
- BBa\_K332011
- BBa\_K284999
- BBa\_K417001
- BBa\_K417002
- BBa\_K653000
- BBa\_K1067002
- BBa\_K1067003
- BBa\_K1152003
- BBa\_K1067004
- BBa\_K1067005
- BBa\_K1067006
- BBa\_K137000
- BBa\_K137014
- BBa\_K137067
- BBa\_K078102
- BBa\_K078003
- BBa\_K078005
- BBa\_K078006
- BBa\_K078007
- BBa\_K078008
- BBa\_K078009
- BBa\_K144000
- BBa\_K156012
- BBa\_K174000
- BBa\_K398000
- BBa\_K398006
- BBa\_K398005
- BBa\_K826001
- BBa\_K808010
- BBa\_K808025
- BBa\_K808026
- BBa\_K863006
- BBa\_K863005
- BBa\_K863000
- BBa\_K863001
- BBa\_K863010
- BBa\_K863011
- BBa\_K863012
- BBa\_K863020
- BBa\_K863015
- BBa\_K863021
- BBa\_K1031620
- BBa\_K1031621
- BBa\_K1031622
- BBa\_K1031623
- BBa\_K1031624
- BBa\_K1031625
- BBa\_K1031931
- BBa\_K1031923
- BBa\_K1031922



## Apéndice 10. Glosario

- *ADN*: abreviación para ácido desoxirribonucleico, contiene las instrucciones genéticas usadas en el desarrollo y funcionamiento de todos los organismos vivos y algunos virus. También es responsable por su transmisión hereditaria.
- *Aminoácidos*: Un aminoácido es una molécula orgánica con un grupo amino (-NH<sub>2</sub>) y un grupo carboxilo (-COOH). Los aminoácidos más frecuentes y de mayor interés son aquellos que forman parte de las proteínas.
- *API (Application Programming Interface)*: conjunto de funciones y procedimientos que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción. Esta especifica cómo algunos componentes de software deben interactuar unos con otros. Una serie de rutinas, protocolos y herramientas para la construcción de aplicaciones de software.
- *CDS (Coding DNA Sequence)*: es una sección del ADN o RNA, compuesta de exones, que codifica la fabricación de una proteína.
- *Cola*: estructura de datos abstracta en donde sus elementos son mantenidos en cierto orden, la adición de nuevos elementos se realizar por el fin y la eliminación mediante el inicio de la misma.
- *Compuestos*: sustancia formada por la unión de dos o más elementos de la tabla periódica.
- *Enzimas*: moléculas que catalizan reacciones químicas.
- *Estructura terciaria*: en general es la figura compleja de las moléculas de una proteína individual. Muestra la relación espacial de una estructura secundaria con otra. Este tipo de estructura controla la función básica de una proteína.
- *Eucariotas*: se denominan células eucariotas todas aquellas que posean su núcleo y otras estructuras rodeadas de una membrana, como el ser humano.
- *Exones*: es la región del gen que no es separada durante el proceso de corte del ARM, por lo que me mantiene en el ARN mensajero maduro. Estos contienen la información para producir la proteína codificada en el gen.

- *Gen*: es una secuencia ordenada de nucleótidos en la molécula de ADN. Además, es la unidad molecular de herencia en un ser vivo.
- *Genoma*: información genética (hereditaria) de un organismo o una especie en particular.
- *Intrones*: región dentro del gen que debe ser eliminada durante la transcripción del ADN.
- *JUNG*: por sus siglas en inglés, Java Universal Network/Graph Framework.
- *KEGG*: por sus siglas del inglés, Kyoto Encyclopedia of Genes and Genomes.
- *Metagenómica*: estudio del conjunto de genomas de un determinado entorno, directamente en su ambiente sin necesidad de aislar y cultivar esas especies.
- *Nucleótidos*: molécula biológica que forma los bloques básicos de los ácidos nucleídos (ADN y el RNA). Sus bases se representan como: A,T,C,G,U.
- *Plásmido*: moléculas de ADN circular o lineal que se replican y transcriben independientes del ADN cromosómico. Presentes en las bacterias.
- *Plásmido*: moléculas de ADN que se replican y transcriben independientes al ADN cromosómico, estos se encuentran en bacterias y algunas levaduras.
- *Procariotas*: se denominan células procariotas, todas aquellas que no posean un núcleo rodeado por una membrana, como las bacterias.
- *Proteínas*: moléculas formadas por cadenas lineales de aminoácidos.
- *Protocolos (de comunicación)*: conjunto de reglas y normas que permite que las entidades en un sistema de comunicación se comuniquen para poder transformar información.
- *REST (REpresentational State Transfer)*: es una arquitectura simple que generalmente corre sobre el protocolo HTTP. REST involucra la lectura de una página web designada que contiene un archivo XML, el cual describe este contenido.
- *Rutas metabólicas*: sucesión de reacciones químicas que conducen de un sustrato inicial a uno o varios productos finales, a través de una serie de metabolitos intermediarios.

- *Scripts:* Un script es un guión o conjunto de instrucciones. Permiten la automatización de tareas creando pequeñas utilidades.
- *Secuenciación:* conjunto de métodos y técnicas bioquímicas con el fin de determinar el orden de los nucleótidos (A,C,G,T) en el ADN.
- *Substratos:* una molécula sobre la que actúa una enzima.
- *Usuario:* profesional con experiencia en biología o biotecnología que utilizará el herramienta diseñada.
- *XML:* por sus siglas en inglés, Extensible Markup Language.