Los modelos metabólicos a escala genómica. Una propuesta algorítmica para su depuración.

Metabolic models at genomic scale. An algorithmic proposal for its purification.

MsC. Hendy Maier Pérez Barrera

DrC. Julian Triana Dopico

DraC. Raymari Reyes Chirino

Universidad de Pinar del Río. Cuba

Universidad Salesiana. Sede Ecuador

Universidad de Pinar del Río. Cuba

Resumen

En las últimas décadas, los avances en la biología molecular y en la infraestructura disponible para el desarrollo de investigaciones en este campo, han dado lugar a una creciente acumulación de datos biológicos. En este escenario, la Biología de Sistema ha surgido como una prometedora ciencia, en esta área, las vías metabólicas y sus capacidades constituyen los objetos de estudios prioritarios.

La reconstrucción de los modelos metabólicos a escala genómica es un proceso que tiene como objetivo permitir la simulación del metabolismo celular del organismo en estudio. La proximidad de las simulaciones al comportamiento in vivo, dependerá de la calidad de los datos biológicos así como de la calidad de estos modelos a gran escala. Este proceso, no automatizado e iterativo, presupone el trabajo a largo plazo de un especialista utilizando la información contenida en diversas bases de datos biológicas.

Se propone fundamentar el proceso de simulación del metabolismo celular de los organismos, basado en métodos estadísticos y ortológicos que permita implementar un algoritmo para la depuración eficiente de los modelos metabólicos reconstruidos a escala genómica.

Para ello, como resultado de esta investigación se diseñó e implementó una herramienta informática que basada en el algoritmo desarrollado garantice la depuración de estos modelos en función de lograr

una mayor exactitud en la simulación del metabolismo celular de los organismos.

El algoritmo desarrollado permitirán acelerar el proceso de reconstrucción de modelos metabólicos a escala genómica a un período de pocos días, dando paso al desarrollo de investigaciones futuras.

Palabras clave: Modelos metabólicos, algoritmo para la depuración, aplicación web.

Abstract:

In recent decades, advances in molecular biology and available for the development of research in this field infrastructure have led to a growing accumulation of biological data. In this scenario, the system Biology has emerged as a promising science, in this area, metabolic pathways and their capacities are the objects of priority studies.

The reconstruction of the genomic scale metabolic models is a process that aims to allow the simulation of cellular metabolism of the organism under study. The proximity of the in vivo behavior simulations depend on the quality of biological data as well as the quality of these models on a large scale. This process is not automated, iterative, presupposes long-term work of a specialist using the information contained in various biological databases.

It is proposed to base the simulation of cellular metabolism of organisms, based on statistical methods and ortológicos implement an algorithm that allows for efficient purification of metabolic models reconstructed genomic scale.

To do this, as a result of this research was designed and implemented a computer-based tool developed algorithm ensures debugging of these models in terms of achieving greater accuracy in the simulation of cellular metabolism of organisms.

The algorithm developed will speed up the process of rebuilding genomic scale metabolic models for a period of a few days, leading to the development of future research.

KeyWords: *Metabolic models, algorithm for debugging web application.*

Introducción

La Biología de Sistemas es el área de investigación que se encarga de estudiar todas las interacciones que se producen dentro de los sistemas biológicos, vistos desde un enfoque sistémico [CITATION JLS06 \l 1033]. Como una evolución conceptual de la Biología de Sistemas surge la Biología Sintética, la cual implica el diseño y construcción de nuevos sistemas biológicos, así como el rediseño de los ya existentes. La misma ha emergido como una poderosa herramienta para la creación de sistemas biológicos noveles, especialmente en el campo de la ingeniería metabólica.

El creciente flujo de la información requiere el uso de técnicas innovadoras para su visualización,

modelación, interpretación y análisis. Las ciencias de la computación, las ciencias de la información, la ingeniería, las matemáticas, la estadística, entre otras han sido aplicadas a la biología, emergiendo así la disciplina hoy conocida como bioinformática.

En este escenario, constituye piedra angular de ésta área, la reconstrucción de los modelos metabólicos a escala genómica. Éste es un proceso que tiene como objetivo permitir la simulación del metabolismo celular del organismo en estudio, permitiendo la integración de información genómica con actividades metabólicas observadas a través de experimentos fenotípicos y otros datos "ómicos" para obtener conocimiento biológico oculto y que pudiera ser de otro modo difícil de obtener [CITATION kFa11 \l 1033]. La proximidad de las simulaciones al comportamiento in vivo, dependerá de la calidad de los datos biológicos así como de la calidad de estos modelos a gran escala.

Se han desarrollado varios métodos que inciden en el refinamiento de la red. El primer método publicado fue SMILEY, que basado en la programación lineal, identifica el mínimo de reacciones que necesitan ser agregadas al modelo metabólico desde una base de datos universal de reacciones, para permitir un acercamiento en el crecimiento que debe ser logrado [CITATION JLR03 \l 1033]. [CITATION VSK07 \l 1033], y colaboradores, proponen dos métodos, GapFind, para identificar los metabolitos no producidos y GapFill, para resolver los huecos con el mínimo de modificaciones en el modelo metabólico.

Sin embargo, no existen algoritmos soportados sobre criterios estadísticos y ortológicos (estudio de homología de secuencias genéticas) que permitan lograr una completitud para dichos modelos. Esta razón hace aún más engorroso el proceso de depuración si se tiene en cuenta que conlleva, según [CITATION Bar14 \l 3082]:

"realizar un estudio de la reversibilidad de las reacciones, metabolitos desconectados y reacciones bloqueadas asociadas a los mismos, así como la inclusión de reacciones que fueron estudiadas para otros organismos y que se puede inferir, mediante enfoques ortológicos, su presencia también para los organismos de estudio" (p.2).

Este desafío incluye, por tanto, la posibilidad de identificar, rellenar o eliminar aquellos metabolitos que no pueden ser producidos por ninguna reacción o importados a través de las vías metabólicas en el modelo y los metabolitos que no son consumidos por ninguna reacción en la red o que no son exportados hacia ninguna ruta existente, así como la eliminación de ciclos internos que no son termodinámicamente factibles, a saber, los ciclos fútiles.

Éstas insuficiencias en la simulación del metabolismo celular de los organismos, a favor de una depuración eficiente de los modelos metabólicos reconstruidos a escala genómica constituye el problema científico a resolver, el justifica el objetivo del presente trabajo de fundamentar el proceso de simulación del metabolismo celular de los organismos, basado en métodos estadísticos y ortológicos que permita implementar un algoritmo para la depuración eficiente de los modelos metabólicos reconstruidos a escala genómica.

Para lograr este objetivo es importante realizar un estudio de los referentes teóricos, así como, un diagnóstico del estado actual del proceso de simulación del metabolismo celular de los organismos. En correspondencia se realizará el diseño e implementación de una herramienta informática que basada en el algoritmo desarrollado, garantice rellenar o eliminar aquellas reacciones que involucran metabolitos

que no pueden ser producidos por ninguna reacción o importados a través de las vías metabólicas en el modelo. Adicionalmente, se pretende simular aquellas reacciones donde los metabolitos que no sean consumidos en la red o no exportados hacia ninguna ruta existente, así como la eliminación de ciclos internos que no son termodinámicamente factibles.

Algunas consideraciones teóricas básicas

Los modelos metabólicos se han utilizado a lo largo de la historia para caracterizar los sistemas biológicos y desarrollar estrategias no intuitivas para una reingeniería de ellos para el aumento de la producción de bioproductos valiosos, [CITATION Ran10 \l 23562]. Más recientemente, los modelos se han desarrollado y aplicado por varios objetivos, por ejemplo: drogas, enfermedades metabólicas, obtención de biocombustibles, el estudio de la patogenicidad microbiana y el parasitismo [CITATION Pit10 \l 23562].

La validación de alta calidad [CITATION Thi10 \l 1033] de los modelos, es fundamental no sólo para una recapitulación de las propiedades fisiológicas conocidas sino también mejorar su precisión de la predicción. Hacia este fin, se han desarrollado estrategias para incorporar otros procesos celulares, tales como: genes, expresión de proteínas. Por ejemplo, los modelos a escala genómica de patógenos se han reconstruido para desarrollar nuevos fármacos en favor de combatir infecciones y también minimizar los efectos secundarios en la acogida [CITATION HUK11 \l 1033].

Los modelos metabólicos pueden facilitar enormemente la evaluación de los posibles fenotipos metabólicos alcanzables por los organismos. Por lo tanto, el rápido desarrollo de alta calidad metabólica, modelos y algoritmos para el análisis de su contenido son de vital importancia. La generación automatizada de los modelos metabólicos a escala genómica, mejoras y aplicaciones, constituyen hoy grandes desafíos en aras de garantizar una rápida reconstrucción del modelo metabólico con una alta calidad.[CITATION IPa12 \l 1033].

Materiales y métodos

Un modelo metabólico contiene una serie de reacciones enzimáticas consecutivas que generan productos específicos. De modo que los productos de una reacción pueden ser los sustratos de otra. De cada reacción se conocen los sustratos, los productos y la enzima que cataliza la misma.

A partir de esta idea se puede modelar el modelo como un grafo dirigido G= (V, A) para su implementación donde V es el conjunto de vértices y A el conjunto de arcos. Los vértices pueden representar las reacciones, los compuestos y las enzimas; y los arcos indican las relaciones que existen entre ellos.

Teniendo en cuenta estos elementos y las siguientes reacciones se puede conformar el grafo de la figura 1.

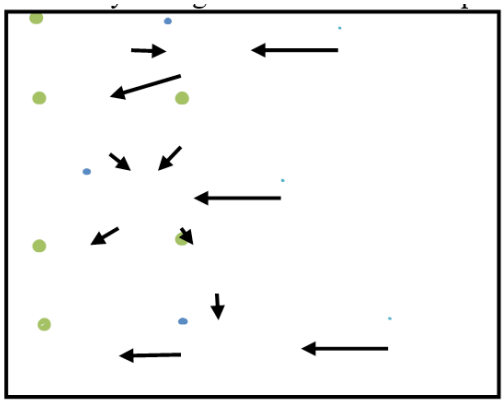

Figura 1. Grafo de reacciones

R1: A=>B, Ec1 **(1)**
R2: B+C=>F+G, Ec3 **(2)**
R4: F=>K, Ec4 **(3)**

donde:
R1, R2 y R4: reacciones
A, B, C, F, G y K: compuestos
Ec1, Ec3, Ec4: enzimas

Para la obtención del modelo metabólico es necesario realizar un análisis en las Bases de datos biológicas disponibles en internet, en las anotaciones del genoma, en los libros de bioquímica; así como en sus más recientes publicaciones. Para luego obtener el modelo a través de una matriz estequiométrica de sus cofactores teniendo en cuenta la reversibilidad de las reacciones, su estequiometria, los cofactores, la ecuación de la biomasa para su posterior predicción.

En este sentido, varios autores han encaminado sus estudios hacia la búsqueda de soluciones informáticas, por ejemplo, Merlin es una aplicación Java de uso fácil que realiza la reconstrucción de los modelos metabólicos a escala genómica para cada organismo que tiene su genoma secuenciado, utilizando herramientas de homología como BLAST y HMMER. Para cada gen, la información de homología se recupera y los resultados se puntúan automáticamente, permitiendo al usuario cambiar la selección automática, y dinámicamente (re-) anotar el genoma.

Además, incluye herramientas para la identificación y anotación de genes que codifican proteínas de transporte, así como la generación de reacciones de transporte para dichos portadores. También se desarrollaron e integraron en Merlin herramientas para la compartimentación del modelo que predicen la localización de las proteínas codificadas en el genoma y, por tanto, la localización de los metabolitos implicados en las reacciones inducidas por dichas proteínas [CITATION Osc15 \l 23562].

En nuestro grupo de desarrollo se ha creado una herramienta informática basada en la web, *Computational Platform to Access Biological Information* (COPABI), donde se logra obtener y reconstruir modelos metabólicos a escala genómica cumpliendo los criterios de completitud y unicidad de las vías metabólicas [CITATION Rai13 \l 1033].

La plataforma produce la red metabólica generada en diferentes salidas: ya sea como un archivo de SBML o directamente como un formato de archivo OptGene que podrían ser analizado directamente en otros softwares o módulos de esta plataforma para su análisis. Como consecuencia se implementará el algoritmo que se quiere desarrollar, agregándole así, nuevas funcionalidades a esta plataforma.

Lo anterior constituye punto de partida esencial para la concepción de la propuesta de solución que se defiende.

Un punto de partida para la reconstrucción automática a escala genómica de los modelos metabólicos es la obtención de la información relevante sobre el organismo para el que el modelo va a ser generado, a saber, la lista de las reacciones , los genes , metabolitos , y enzimas presentes en la célula estudiada. Esta información está disponible a partir de las bases de datos de libre acceso públicos. Sin embargo, la falta de calidad en algunas entradas de las bases de datos son un inconveniente que se debe modificar: los falsos positivos, falsos negativos, así como los objetos anotados erróneamente, puede plantear obstáculos en los esfuerzos para recopilar una lista correcta significativa de las reacciones. Como consecuencia, la reconstrucción debe hacerse bajo estricto control de todas las reacciones, la ecuación de la biomasa debe basarse en moléculas constituyentes, y la coherencia y la integridad de la red deben ser requisitos previos para la generación de un modelo de calidad y útil. [CITATION AMF091 \l 1033].

Obtención y conservación de la información biológica

Una vez que se tiene definida la base de datos se hace necesario obtener la información biológica que se almacenará en la misma. En la actualidad existen gran cantidad de bases de datos de información biológica de acceso público en la red, entre las que destacan las siguientes:

- **BRENDA**: es una base de datos de enzimas que contiene información tanto de las enzimas como de las reacciones enzimáticas y se encuentra disponible en http://www.brenda-enzymes.info . [CITATION ACh09 \l 1033]

- **Biocyc**: disponible en http://biocyc.org/, constituye una colección de bases de datos que suministra diferentes fuentes de referencia de las vías metabólicas de distintos organismos.

- **KEGG**: se encuentra disponible en http://www.kegg.jp/kegg/. La misma está compuesta por elementos genéticos (KEGG GENES), compuestos químicos de sustancias tanto endógenas como exógenas (KEGG LIGAND), redes de reacciones e interacciones moleculares (KEGG PATHWAY), así como jerarquías y relaciones entre varios elementos biológicos (KEGG BRITE). [CITATION Kan10 \l 1033].

Específicamente se trabajó con KEGG, una base de datos de sistemas biológicos, que se nutre de muchas de las otras bases de datos existentes y proporciona valiosos medios para acceder a la información como son: un servicio FTP (KEGG FTP), un servicio web (KEGG API), servicio de Ontología Genética "KEGG Orthology (KO) System".

Teniendo en cuenta que el servicio web no permite realizar descargas masivas se descargó la información a través del servicio FTP. El volumen de información obtenida se encuentra en el orden de las 10000 reacciones, más de 17000 compuestos y alrededor de 6500 enzimas. Es por ello que fue necesario el

desarrollo de una herramienta informática que posibilitara almacenar de forma automática la información biológica en una base de datos.

La aplicación se implementó usando CodeIgniter, framework para construir aplicaciones web usando PHP, basado en el patrón de desarrollo Modelo-Vista-Controlador. Ésta genera una serie de salidas importantes para que el especialista pueda hacer sus respectivos análisis, por ejemplo: en cuanto al listado reacciones, el número de metabolitos que intervienen en el modelo, las posibles reacciones bloqueadas, los metabolitos finales de cadenas, y los posibles ciclos fútiles o ciclos de sustratos, entre otro como se muestra en la figura 2.

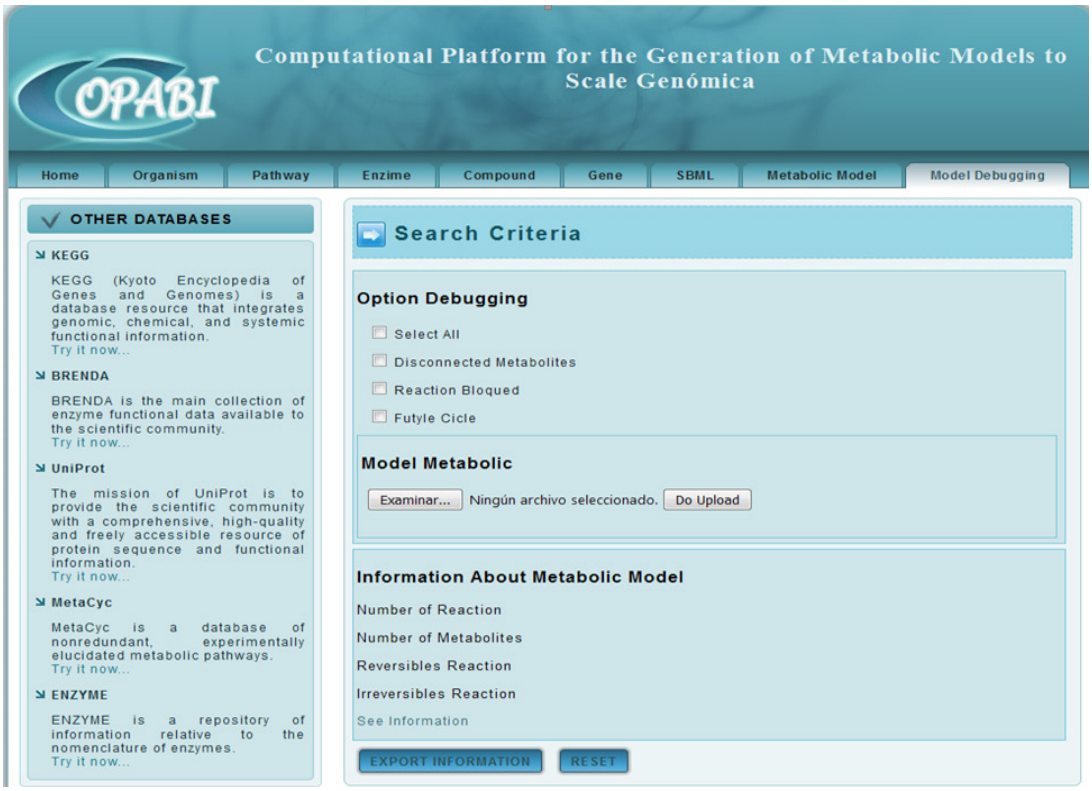

Figura 2. Diseño de la Aplicación Web.

De igual manera se generan estas salidas para otros modelos que estén en formato OptGene o SBML, no necesariamente generados por COPABY.

Un aspecto importante en las salidas de la aplicación, es la correspondiente a minimizar la cantidad de huecos que existen en el modelo metabólico en cuestión, para ello, se modeló e implementó una función objetivo que resuelva este problema, donde se describe a continuación.

$$\text{Minimize } Z = a \cdot V$$
$$\text{subject to}$$
$$S \cdot v = 0$$
$$v_{i,irr} \in \Re^{+}$$
$$v_{i,rev} \in \Re,$$

Z función objetivo
S se define como la matriz estequiométrica con sus coeficientes
i metabolito en las reacciones.
V es el vector de flujo que corresponde al flujo de la reacción j-ésimo.

Figura 3. Función objetivo

Procedimiento para el relleno de los "gaps" usando ontología de genes

En primer lugar la aplicación guarda en un archivo de texto las reacciones que están "antes" y "después" de los "gaps". Puede pasar que no existan las reacciones "antes" o "después" (o sea, que no exista la arista que una a ese nodo con otro, ya sea como producto o como reaccionante). Por ejemplo:

"gap"

Aquí hay una reacción antes y una después del "gap"

"gap"

Aquí no hay reacciones después del "gap"

"gap"

Aquí no hay reacciones antes del "gap"

En el archivo se guardan los nombres de las reacciones, usando los números EC de las enzimas, nombres de las reacciones según los ID usados en KEGG.

Se puede guardar la información pertinente al ID de la reacción, los metabolitos involucrados, etc.

Se obtiene información de la base de datos KEGG, donde se destaca la relación que existe entre las categorías y sub-categorías de organismos y la ontología de genes.

Al rellenar un "gap" se tiene en cuenta las reacciones/genes que están presentes en otros organismos emparentados filogenéticamente (o sea que estén dentro de una misma categoría y sub-categoría). Por ejemplo: el hombre está emparentado con el chimpancé, la gallina con la pata, etc. Cuánto más cercano está un organismo de otro, más probabilidad tendrá de presentar genes comunes, y por tanto de tener la reacción que le falta ("gap").

De esta forma la aplicación informática brinda la opción al usuario de las posibles reacciones/genes que se puede incorporar en un modelo metabólico incompleto, o sea con "gaps". Incluso se brinda un cierto estimado probabilístico de que esté una reacción/gen el modelo metabólico de estudio. Finalmente, se representa el modelo/grafo con los "gaps" corregidos y se le muestra al usuario.

Resultados y discusión

Como resultado de esta investigación se desarrolló una aplicación Web: COPABY V1.1, que permite además de sus funciones ya implementadas, cargar modelos metabólicos a escala genómica y generar todas las posibles incongruencias que existan para su posterior depurado.

Esta depuración se realiza basada en primer lugar en métodos estadísticos siguiendo criterios de unicidad, propuestos COPABY V1.0, que aún no se han descritos en esta investigación, y en segundo lugar en términos de ontología de genes ya antes descrito.

Los modelos, según las salidas de la aplicación realizada, oscilan entre los 1024 y los 1100 vértices (reacciones, los compuestos y las enzimas) y más de 900 arcos, permitiendo obtener diferentes consultas relacionadas con estos modelos, cómo se muestra a en la figura 3.

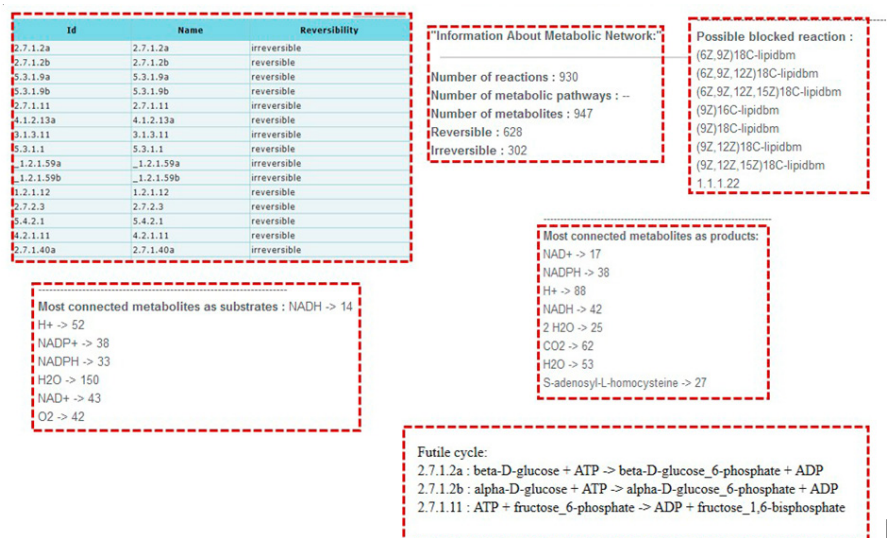

Figura 3. Algunas salidas de la aplicación

En una primera aproximación a la evaluación teórica de la propuesta, se realizó una Consulta a Especialistas sobre el tema, quienes coinciden en su pertinencia a favor de la solución del problema identificado.

Conclusiones

Los diferentes aspectos que se abordan en la investigación, el problema planteado, los objetivos propuestos y el análisis realizado permiten arribar a las siguientes conclusiones:

- El estudio de los referentes teóricos abordados en la presente investigación contribuyó a una mejor organización, desde el punto de vista de la depuración de los modelos metabólicos y de su proceso de ingeniería.

- El diagnóstico del estado actual en el proceso de depuración de los modelos metabólicos permitió, de manera coherente, articular los procesos computacionales (conservar, procesar, recuperar) como contenido de la aplicación que se realizó.

- El diseño e implementación de la aplicación Web que se propone, significa una alternativa para optimizar el proceso que demanda el cliente, mediante una interfaz amigable y sencilla.

Referencias

Barrera, H. M. (2014). Un acercamiento a la depuración de modelos metabólicos a escala genómica. *Revista Mendive*.

Chang, A., Scheer, M., & Grote, A. (2009). BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids*.

Dias O, M. R. (2015). Reconstructing genome-scale metabolic models with merlin. *Oxford Journals, Science & Mathematics, Nucleic Acids Research, Volume 43, Issue 8.*, 3899-3910.

Fang, k., Zhao, H., Sun, C., Lam, C., Chang, S., & Zhang, K. (2011). Exploring the metabolic network of the epidemic pathogen Burkholderia cenocepacia J2315 via genome-scale reconstruction. *Systems biology*.

Feist, A., Herrgard, M., & Thiele, I. (2009). Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*, 129–143.

Kanehisa, M., & Goto., S. (2010). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27-30.

Kim, H., Kim, S., & Jeong, H. (2011). Integrative genome-scale metabolic analysis. *Mol Syst*.

Kumar, V., Dasika, M., & Maranas, C. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* .

Pagani, I., Liolios, K., Jansson, J., & Chen. (2012). The Genomes OnLine Database. *Nucleic Acids*.

Pitkanen E, R. J. (2010). Computational methods for metabolic reconstruction. . *Curr Opin Biotechnol* .

Ranganathan S, S. P. (2010). OptForce: an optimization procedure for identifying all genetic. *PLoS Comput Biol*.

Reed, J., Vo, T., Schilling, C., & Palsson, B. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). . *Genome biology*.

Reyes, R. (2013). DESARROLLO Y ANÁLISIS DE ALGORITMOS PROBABILÍSTICOS PARA LA RECONSTRUCCIÓN DE MODELOS METABÓLICOS A ESCALA GENÓMICA.

Snoep, J., Bruggeman, F., & Olivier, B. (2006). Towards building the silicon cell: A modular approach. . *BioSystems*, 207–216.

Thiele, I., & Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 93–121.