UNIVERSIDAD POLITÉCNICA DE VALENCIA
DEPARTAMENTO DE MATEMÁTICA APLICADA

DESARROLLO Y ANÁLISIS DE ALGORITMOS PROBABILÍSTICOS PARA LA RECONSTRUCCIÓN DE MODELOS METABÓLICOS A ESCALA GENÓMICA

TESIS DOCTORAL

Presentada por:
Raymari Reyes Chirino

Dirigida por:
Pedro José Fernández de Córdoba Castellá
Javier Fermín Urchueguía Schölzel

Valencia, 2013

Frase

No basta dar pasos que un día puedan conducir hasta la meta, sino que cada paso ha de ser una meta…

Johann P. Eckermann

Dedicatoria

A mis preciosas Carolina y Sabrina, porque este resultado es por ustedes y para ustedes.

Agradecimientos

Este resultado constituye el más grande de mis metas como profesional. Durante este tiempo de investigación han pasado muchas cosas en mi vida, unas planificadas y otras no, pero que nunca hicieron que dejara de luchar por este resultado. Por el contrario, me dieron la fuerza necesaria para seguir adelante con este sueño.

Son muchas las personas a las cuales debo agradecer por su apoyo durante esta etapa.

- En primer lugar a mis directores de tesis, Pedro Fernández de Córdoba y Javier Fermín Urchueguía Schölzel por confiar en nuestro grupo desde el primer momento. Apenas un correo y nuestro interés por investigar, hicieron que rápidamente se pensara en los informáticos para comenzar una nueva línea de investigación dentro del consolidado grupo de InterTech. Ellos, con esa sabiduría que les caracteriza, buscaron puntos en común para adentrarnos en el fascinante mundo de la Biología de Sistemas. Gracias Pedro y Javi por permitirme formar parte de la gran familia que es InterTech. Con ustedes aprendí lo que es ser ejemplo de profesional y ser humano a la misma vez.
- A todo el grupo de InterTech. Y los incluyo a todos, los que estuvieron y los que están. Gracias por formar esa linda familia donde la ciencia se conjuga con cariño y una bella amistad.
- A Malena, María Amparo y Betty, por darnos la luz para emprender este camino.
- A mis queridos compañeros del Departamento de Informática, sin ustedes no hubiese podido lograr este resultado. Su apoyo y comprensión en cada momento me permitieron concentrarme por completo en esta dura tarea. Cary por ser mi madre en la Universidad y Albe por permitirme que me *“robe”* su espacio de trabajo. A todos gracias.
- A todas las personas de la Universidad y la Facultad que me brindaron su apoyo para ayudarme en cada una las gestiones, que ya saben que no son pocas. No mencionaré nombres porque son unos cuantos, y no quiero correr el riesgo de olvidar a alguien. Ellos saben quiénes son.

- A Yarlenis, Javier, Leslie y Dariel, por la seriedad con que asumieron cada una de las tareas y por la alta responsabilidad demostrada durante sus trabajos de diploma.
- A Julián, para ti un agradecimiento especial. Porque sacrificas tu tiempo para ayudarme. Porque para ti no hay hora ni momento inapropiado cuando se trata de trabajar. Eres un excelente profesional.
- A Daniel, por tu ayuda durante mi segunda estancia en Valencia. Por inculcarme tu ejemplo de gran investigador. Por darme fuerzas y hacerme entender que todo sale, aún cuando la aplicación tardaba días en descargar un organismo.
- A mis amigos en Valencia. A Mayrelis, Víctor y su familia, por acogerme en su casa como si fuera parte de ellos, por hacer que mis días en Valencia fueran más amenos. A Antonio y Carmen, porque cuidaron de mi como una nieta.
- A mima y papi, porque el mundo sabe lo que significan para mí. Por su entrega a Carolina como si fuera yo y por apoyarme siempre en cada uno de mis empeños.
- A mami, por estar siempre a mi lado aunque nos separen 1500 kilómetros de distancia. Para nosotras no hay barreras. Gracias por entenderme siempre.
- A Dargenky y su familia, porque han estado en cada momento crucial de mi vida. A ti mi amor porque te conocí justo cuando más te necesitaba. Por ser mi compañero, mi confidente y mi amor. Por los momentos lindos que hacen más fuerte nuestro amor.
- A mi papá, que ha sabido sacrificarse por nosotros y ha enfrentado momentos difíciles. Por ser ejemplo de ser humano y de profesional. Gracias mi amor porque nunca me ha faltado tu consejo.
- A mis tres hermanas, porque son bellas por dentro y por fuera. Porque somos como una sola aunque estemos separadas.
- A toda la familia, por quererme tanto y estar siempre a mi lado.
- A mis amigos, por estar siempre ahí.

A todos muchas gracias!!!!

Resumen

Los avances en la Biología Molecular y las técnicas genómicas, las nuevas herramientas bioinformáticas que han posibilitado el acceso a miles de datos biológicos, el mayor poder computacional y los nuevos algoritmos de modelación han propiciado el nacimiento de una nueva disciplina denominada Biología de Sistemas. Esta nueva área de investigación se centra en el estudio de un sistema biológico, analizado como un sistema integrado de biomoléculas y reacciones bioquímicas interrelacionadas que dan lugar a la gran variedad de procesos biológicos. Profesionales procedentes de áreas como la Biología, la Informática, la Física, la Química o las Matemáticas se han conjugado en el estudio de esta ciencia moderna.

Uno de los enfoques fundamentales de dicha rama se basa en la reconstrucción de modelos metabólicos a escala genómica, un esfuerzo que a día de hoy no se encuentra automatizado. Dicho proceso consiste en listar y agrupar el conjunto de reacciones metabólicas de un organismo, a partir de la información disponible en diversas bases de datos biológicas, por lo que requiere del trabajo de un especialista durante varios meses. Los modelos metabólicos a escala genómica constituyen una herramienta útil para estudiar las capacidades metabólicas de un organismo y su comportamiento ante posibles perturbaciones, con lo cual es posible diseñar estrategias ingenieriles orientadas a mejorar una función en particular.

El presente proyecto se basó en el desarrollo y análisis de algoritmos que incluyen decisiones a partir de criterios probabilísticos. Consecuentemente, se logra reconstruir modelos metabólicos a escala genómica cumpliendo los criterios de completitud y unicidad de las vías metabólicas. Como parte del algoritmo se trató la inclusión de reacciones metabólicas adicionales al modelo. Su selección se fundamentó por la prevalencia de metabolitos en un mapa metabólico general, conformado por todas las reacciones metabólicas que existen en los sistemas vivos de la naturaleza. Por otro lado, se tuvo en cuenta la presencia repetida de una misma reacción metabólica pero relacionada con diferentes enzimas. Nuevamente, se usó un criterio probabilístico para

la toma de decisión basada en la unicidad de las vías metabólicas, considerando una única reacción bioquímica.

La metodología seguida en la automatización de este proceso fue la implementada de forma manual para la reconstrucción del primer modelo metabólico a escala genómica de un microorganismo fotosintético, la *Synechocystis sp. PCC6803*. Como resultado se obtuvo además, la aplicación web *Computational Platform to Access Biological Information* (COPABI) que permite reconstruir modelos metabólicos a escala genómica siguiendo la metodología antes mencionada. Para la validación de los resultados se compararon 9 modelos metabólicos de organismos publicados en la literatura con los modelos generados por COPABI siguiendo los criterios probabilísticos al 10 % y al 100%. Se midieron indicadores como: cantidad de reacciones, cantidad de metabolitos, cantidad de pares de metabolitos conectados entre sí, porcentaje de reacciones reversibles e irreversibles, entre otros. Además, se utilizaron algoritmos estándar que permitieron calcular el promedio de la ruta más corta entre los nodos de la red y la conectividad de los mismos. Los resultados de la comparación entre los dos modelos automáticos generados por COPABI y el modelo utilizado de la literatura muestran la tendencia de la distribución siguiendo una ley de potencia y la similitud entre los modelos. Una vez más se demuestra la efectividad de la metodología implementada para la reconstrucción de modelos metabólicos a escala genómica. Por último, se procedió a evaluar la semejanza entre dos modelos metabólicos a partir de un criterio que permite diferenciar ambas redes. Los resultados obtenidos de la comparación de 6 modelos metabólicos de diferentes organismos, demuestran la consistencia de los modelos reconstruidos automáticamente.

Los algoritmos probabilísticos desarrollados 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.

Abstract

The advances in molecular biology and genomic techniques, and the new bioinformatics tools that have enable access to thousands of biological data, the greater computational power and the new modeling algorithms have propitiated the appearance of a new discipline called System Biology. This new research area is focused on the study of a biological system analyzed as an integrated system of biomolecules and interrelated biochemical reactions that lead to the great variety of biological processes. Professionals from fields such as: Biology, Computer Science, Physics, Chemistry and Mathematics have been joined together to study this modern science.

One of the fundamental approaches of these fields is based on the reconstruction of genome-scale metabolic models, effort that today has not been automated. This process consist on listing and grouping the set of metabolic reactions of an organism, from the information available in different biological database, therefore it is needed the labor of a specialist for a months. The genome-scale metabolic models constitute a useful tool to study the metabolic capabilities of an organism and its behavior front possible disturbances, which makes possible to design engineering strategies aimed to improve a particular function.

This project focused on the development and analysis of algorithms including decisions from probabilistic criteria. Consequently genomic-scale metabolic models can be reconstructed fulfilling the criteria of completeness of metabolic and uniqueness of metabolic pathways. As part of the algorithm it is discussed to include the additional metabolic reactions to the model. Their selection was based on the prevalence of metabolic reactions that appear in the alive systems in nature. Moreover, the presence of the same repeated metabolic reaction was considered, but related with different metabolic enzymes. Again a probabilistic approach was used to take the decision based on the uniqueness of the metabolic pathways considering unique biochemical reactions.

The methodology used in the automation of this process was implemented manually for the reconstruction of the first genome-scale metabolic model of the photosynthetic

microorganism, the Synechocystis sp. PCC6803. As a result, it was also obtained the Computational Platform to Access a Biological Information (COPABI) that can reconstruct genome-scale metabolic models following the above methodology. For validation of the results, 9 metabolic models of organisms published in literature were compared to the generated models by COPABI, following probabilistic approaches to 10% and to 100%, were compared. Indicators such as: quantity of reactions, quantity of metabolites, quantity of pairs of interconnected metabolites, percentage of reversible and irreversible reactions, among others. Standard algorithms were also used, that allowed the calculation of the average shortest path between the network and connectivity of them. The results of the comparison between the generated two COPABI automatic models and the literature model used show the distribution tendency following a power law and the similarity between the models. Once more, the effectiveness of the methodology used for the reconstruction of genome-scale metabolic models is shown. Finally, it was evaluated the similarity between two metabolic models, from a criterion that differentiates the two networks. The results of the comparison of 6 different organisms metabolic models show the consistency of the models automatically reconstructed.

The developed probabilistic algorithm may accelerate the reconstruction process of genome-scale metabolic models to a period of few days, giving the possibility of future researches.

Resum

Els avanços en la Biologia Molecular i les tècniques genòmiques, les noves ferramentes bioinformàtiques que han possibilitat l'accés a milers de dades biològiques, el major poder computacional i els nous algoritmes de modelació han propiciat el naixement d'una nova disciplina denominada Biologia de Sistemes. Esta nova àrea d'investigació se centra en l'estudi d'un sistema biològic, analitzat com un sistema integrat de biomoléculas i reaccions bioquímiques interrelacionades que donen lloc a la gran varietat de processos biològics. Professionals procedents d'àrees com la Biologia, la Informàtica, la Física, la Química o les Matemàtiques s'han conjugat en l'estudi d'esta ciència moderna.

Un dels enfocaments fonamentals d'aquesta branca es basa en la reconstrucció de models metabòlics a escala genòmica, un esforç que a hores d'ara no es troba automatitzat. Aquest procés consistix a llistar i agrupar el conjunt de reaccions metabòliques d'un organisme, a partir de la informació disponible en diverses bases de dades biològiques, per la qual cosa requerix del treball d'un especialista durant diversos mesos. Els models metabòlics a escala genòmica constituïxen una ferramenta útil per a estudiar les capacitats metabòliques d'un organisme i el seu comportament davant de possibles pertorbacions, amb la qual cosa és possible dissenyar estratègies ingenieriles orientades a millorar una funció en particular.

El present projecte es va basar en el desenrotllament i anàlisi d'algoritmes que inclouen decisions a partir de criteris probabilístics. Conseqüentment, s'aconseguix reconstruir models metabòlics a escala genòmica complint els criteris de completesa i unicitat de les vies metabòliques. Com a part de l'algoritme es va tractar la inclusió de reaccions metabòliques addicionals al model. La seua selecció es va fonamentar per la prevalença de metabòlits en un mapa metabòlic general, conformat per totes les reaccions metabòliques que existixen en els sistemes vius de la naturalesa. D'altra banda, es va tindre en compte la presència repetida d'una mateixa reacció metabòlica però relacionada amb diferents enzims. Novament, es va usar un criteri probabilístic per a la presa de decisió basada en la unicitat de les vies metabòliques, considerant una única reacció bioquímica.

La metodologia seguida en l'automatització d'este procés va ser la implementada de forma manual per a la reconstrucció del primer model metabòlic a escala genòmica d'un microorganisme fotosintètic, la *Synechocystis sp. PCC6803*. Com resultat es va obtindre a més, l'aplicació web *Computacional Platform to Access Biological Information* (COPABI) que permet reconstruir models metabòlics a escala genòmica seguint la metodologia abans esmentada. Per a la validació dels resultats es van comparar 10 models metabòlics d'organismes publicats en la literatura amb els models generats per COPABI seguint els criteris probabilístics al 10 % i al 100%. Es van mesurar indicadors com: quantitat de reaccions, quantitat de metabòlits, quantitat de parells de metabòlits connectats entre si, percentatge de reaccions reversibles i irreversibles, entre altres. A més, es van utilitzar algoritmes estàndard que van permetre calcular la mitjana de la ruta més curta entre els nodes de la xarxa i la connectivitat dels mateixos. Els resultats de la comparació entre els dos models automàtics generats per COPABI i el model utilitzat de la literatura mostren la tendència de la distribució seguint una llei de potència i la similitud entre els models. Una vegada més es demostra l'efectivitat de la metodologia implementada per a la reconstrucció de models metabòlics a escala genòmica. Finalment, es va procedir a avaluar la semblança entre dos models metabòlics a partir d'un criteri que permet diferenciar les dues xarxes. Els resultats obtinguts de la comparació de 6 models metabòlics de diferents organismes, demostren la consistència dels models reconstruïts automàticament.

Els algoritmes probabilístics desenrotllats permetran accelerar el procés de reconstrucció de models metabòlics a escala genòmica a un període de pocs dies, donant pas al desenrotllament d'investigacions futures.

Prefacio

Esta tesis doctoral fue desarrollada dentro del marco de colaboración entre el Grupo de Modelización Interdisciplinar, InterTech (www.intertech.upv.es) de la Universidad Politécnica de Valencia y la Universidad de Pinar del Río. Estos lazos se mantienen durante más de 15 años con la institución cubana. Conjuntamente, se ha trabajado con la Cátedra Energesis de la Universidad Católica de Valencia, mediante el intercambio con especialistas en la rama de la Biología de Sistemas.

Las contribuciones originales del presente trabajo de investigación se recogen en los siguientes trabajos:

1. Reyes R., Gamermann D., Montagud A., Fuentes D., Triana J., Fernández de Córdoba P., Urchueguía J. 2012. Automation on the generation of genome scale metabolic models. Journal of Computational Biology, December 2012, 19(12): 1295-1306. doi:10.1089/cmb.2012.0183.

2. Pacheco, Y., Reyes R., Triana, J. 2012. Servicio Web Cliente Orientado a la obtención de la información biológica disponible en la Base de Datos KEGG. Media: Paperback Book, 88 pages. Editorial: Editorial Académica Española. Publication Date: Apr. 6th, 2012. ISBN-10: 3848471051. ISBN-13: 9783848471058.

3. R. Reyes, J. Garrido, R.A. Jaime, V. Córdova, J. Triana, L. Villar, J.C. Castro, P. Fernández de Córdoba, J.F. Urchueguía, E. Navarro y A. Montagud.2011. Desarrollo de una plataforma computacional para el modelado metabólico de microorganismos. *Nereis. Revista Iberoamericana de Métodos, Modelización y Simulación Interdisciplinar.* Universidad Católica de Valencia "San Vicente Mártir". 3: 25-31. ISSN 1888-8550. 2011.

4. Reyes R., Pacheco, Y., Triana J., Gamermann D., Montagud A., Fernández de Córdoba P., Urchueguía J.Integrated database for metabolic models reconstruction using COPABI. En preparación.

Actas de congresos

1. Phylogenic tree reconstruction from metabolic models through the Kruskal algorithm. Gamermann D, Montagud A, Conejero A, Reyes R, Fuente D, de Córdoba PF, Urchueguía J. ECCB´12. *11th European Conference on Computational Biology*. 9-12 September 2012, Basel, Switzerland. Publicado en: F1000 Posters 2012, 3: 1342. Disponible en la siguiente dirección electrónica: http://f1000.com/posters/browse/summary/1092528

2. R. Reyes, R. Jaime, J. Garrido, J. Triana, L. Villar, V. Córdova, J.C. Castro, E. Navarro, A. Montagud, P. Fernández de Córdoba, J.F. Urchueguía y J. Martínez. 2010. Diseño de bases de datos biológicas, un paso hacia la automatización del proceso de construcción de modelos a escala genómica. *XV Convención Científica de Ingeniería y Arquitectura (CCIA 15*). La Habana, Cuba, del 29 de noviembre al 3 de diciembre de 2010.Publicado en: Memorias de la Conferencia. ISBN 978-959-261-317-1.

3. J. Garrido, J. Triana, L. Villar, R. Jaime, R. Reyes, V. Córdova, J.C. Castro, E. Navarro, A. Montagud, P. Fernández de Córdoba y J.F. Urchueguía. 2010. Rational Organism Network Painter: una herramienta optimizada de visualización de redes metabólicas de fácil uso. *XV Convención Científica de Ingeniería y Arquitectura (CCIA 15)*. La Habana, Cuba, del 29 de noviembre al 3 de diciembre de 2010. Publicado en: Memorias de la Conferencia. ISBN 978-959-261-317-1.

4. J. Triana, V. Córdova, R. Jaime, R. Reyes, J. Garrido, L. Villar, F. Márquez, J.C. Castro, E. Navarro, A. Montagud, P. Fernández de Córdoba y J.F. Urchueguía. 2010. Modelo metabólico de una cianobacteria, una fuente de energía a partir de la luz. *I Congreso Internacional de Ingeniería Química, Biotecnológica y Alimentaria (CIIQBA 2010)*. La Habana, Cuba, del 29 de noviembre al 3 de diciembre de 2010. Publicado en: Memorias de la Conferencia. ISBN 978-959-261-317-1.

5. J. Garrido, L. Villar, R. Reyes, R. Jaime, J. Triana, V. Córdova, J.C. Castro, E. Navarro, A. Montagud, P. Fernández de Córdoba, J.F. Urchueguía y J. Martínez. 2011. HYDRA: una plataforma informática orientada al diseño,

análisis y visualización de redes metabólicas. *XIV Convención y Feria Internacional Informática 2011*. La Habana, Cuba, del 7 al 11 de febrero de 2011. Publicado en: Programa Científico. Pág. 55.ISBN 978-959-7213-01-7.

6. R. Jaime, J. Garrido, R. Reyes, L. Villar, J. Triana, V. Córdova, J.C. Castro, E. Navarro, A. Montagud, P. Fernández de Córdoba, J. Urchueguía, J. Martínez y Z. Hernández. 2011. Nueva herramienta para el análisis del balance de flujo de las rutas metabólicas y su integración en Ron Painter. *XIV Convención y Feria Internacional Informática 2011*. La Habana, Cuba, del 7 al 11 de febrero de 2011. Publicado en: Programa Científico. Pág. 153. ISBN 978-959-7213-01-7.

7. R. Reyes, R. Jaime, J. Garrido, J. Triana, L. Villar, V. Córdova, J.C. Castro, E. Navarro, A. Montagud, P. Fernández de Córdoba, J.F. Urchueguía y J. Martínez. 2011. Base de datos biológica orientada a la automatización del proceso de construcción de modelos a escala genómica. Resultados en la Synechocystis SP PCC6803. *XIV Convención y Feria Internacional Informática 2011*. La Habana, Cuba, del 7 al 11 de febrero de 2011.Publicado en: Programa Científico. Pág. 215. ISBN 978-959-7213-01-7.

Tabla de contenidos

Índice de Tablas

Índice de Figuras

Abreviaturas

ADN	Ácido desoxirribonucleico
ARN	Ácido ribonucleico
BS	Biología Sintética
COPABI	*Computational Platform to Access Biological Information*
FQ	Fibrosis Quística
FBA	*Flux Balance Analysis*
MOMA	*Minimization of Metabolic Adjustments*
MFA	*Metabolic Flux Analysis*
EC	*Enzyme Commission*
BM	Ecuación de Biomasa
SWC	Servicio Web Cliente
SGBD	Sistema Gestor de Base de datos
SQL	*Structured Query Language*
IUBMB	*Nomenclature Committee of the International Union of Biochemistry*
HPRD	*Protein Reference Database*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
WWW	*World Wide Web*
SOC	*Service Oriented Computing*
SOA	*Service Oriented Architecture*

WS	*Web Service*
XML	*Extensible Markup Language*
HTML	*HyperText Markup Language*
SOAP	*Simple Object Access Protocol*
HTTP	*Hypertext Transfer Protocol*
WSDL	*Web Services Description Language*
JVM	*Java Virtual Machine*
API	*Application Programming Interface*
IDE	*Integrated Development Environment*
MVCC	*Multiversion Concurrency Control*
ACID	*Atomicity, Consistency, Isolation, Durability*
ANSI	*American National Standards Institute*
CASE	*Computer Aided Software Engineering*
MVC	Modelo-Vista-Controlador
SBML	*Systems Biology Markup Language*
BLAST	*Basic Local Alignment Search Tool*
KGML	*KEGG Markup Language*
UML	*Unified Modeling Language*
SYN	*Synechocystis sp PCC6803*
SYF	*Synechococcus elongatus sp. PCC7942*
BCJ	*Burkhoderia cenocepacia J2315*
RSP	*Sphaeroides Rhodobacter*
CBE	*Clostridium beijerinckii*

MGE	*Mycoplasma genitalium*
LPL	*Lactobacillus plantarum*
TMA	*Thermotoga maritima*
YPK	*Yerisinia pestis*
IUPAC	*Union of Pure and Applied Chemistry*
NC-IUBMB	*Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*
NCBI	*National Center for Biotechnology Information*
PHP	*Hypertext Preprocessor*
RAD	*Rapid Application Development*
WAF	*Web Application Framework*
MILP	*Mixed Integer Linear Programming*
IdentiCS	*Identification of Coding Sequences from Unfinished Genome Sequences*
MrBac	*Metabolic network Reconstructions for Bacteria*
AUTOGRAPH	*AUtomatic Transfer by Orthology of Gene Reaction Associations for Pathway Heuristics*
metaSHARK	*metabolic SearcH And Reconstruction Kit*
GEMSiRV	*GEnome-scale Metabolic model Simulation, Reconstruction and Visualization*
HYDRA	*HYrbid Draw and Routes Analysis*

Introducción

La Biología Molecular de las últimas décadas se ha basado en la teoría que asume el camino directo existente entre genes, proteínas y función biológica, así como la presencia de respuestas predeterminadas del sistema a perturbaciones externas. Aunque este tipo de investigación ha dado lugar a gran cantidad de conocimiento, no proporciona información acerca de cómo integran las células estos datos de forma que se genere un tipo de respuesta u otro. A pesar de que la Biología de Sistemas se considera una nueva disciplina, el estudio de los procesos biológicos como sistemas se trató por Wiener en 1948 en lo que se llamó en aquel momento la cibernética. La Biología de Sistemas ha sido descrita por investigadores como Leroy Hood con detalle, aunque el término ya se empleó por primera vez en 1968 por teóricos como Mesarovic [López et al. (2007)].

I.1 Biología de Sistemas

La Biología de Sistemas es el campo de investigación interdisciplinario de los procesos biológicos en el que las interacciones de los elementos, internos y externos, que influyen en el desarrollo del proceso se representan mediante un sistema matemático. Este enfoque global permite comprender el funcionamiento de los sistemas biológicos y profundizar en el entendimiento de cómo sus interacciones internas y externas (con otros sistemas) conlleva a la aparición de nuevas propiedades y/o procesos.

Convencionalmente, en el estudio de los procesos biológicos se utiliza el método científico clásico, que se basa en la confirmación o refutación de una hipótesis al confrontarla con los resultados experimentales. La Biología sistémica utiliza un enfoque basado en la modelización matemática de los procesos en estudio. Como resultado de la simulación, al poner a funcionar los modelos matemáticos con los que se representa el proceso, se obtiene una serie de predicciones del estado de dicho proceso biológico que corresponderían a los resultados experimentales esperados. Durante las simulaciones, la red de interacciones entre los elementos que componen el proceso biológico se representa mediante un sistema de ecuaciones diferenciales. Los valores de las características de dichos elementos a distintos tiempos y bajo

diversas condiciones experimentales (simuladas), son predecibles debido a que la dinámica del estado de ese sistema modelado es calculable matemáticamente. La Biología de Sistemas es un área interdisciplinaria en la que participan informáticos, biólogos, químicos, matemáticos, físicos, entre otros.

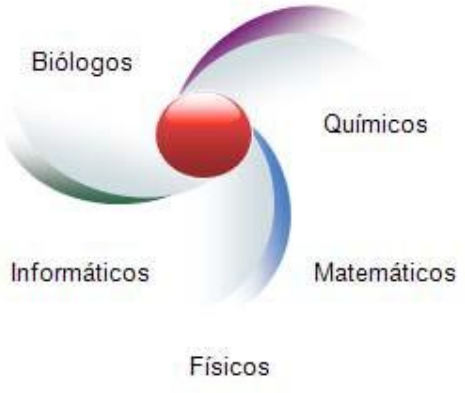

Figura 1. Integración de diversas disciplinas.

Esta disciplina emplea una estrategia diferente a las aproximaciones empíricas tradicionales, por medio del estudio de sistemas biológicos en sus diferentes niveles, desde células y redes celulares hasta organismos completos. La Biología de Sistemas implica el *mapeo* de rutas, interacción de proteínas y genes, así como el de los circuitos de organismos a nivel celular y de organismo completo, todo ello integrado en un modelo informático.

Principales características de la Biología de Sistemas [López et al. (2007)]:

- Estudia los sistemas biológicos de una forma global, a nivel molecular.
- Contrasta con la aproximación clásica linear (un gen, una proteína).
- Integra el conocimiento de diferentes plataformas o disciplinas (genómica, transcriptómica, proteómica, metabolómica, fisiología, patología, etc.).
- Maneja una gran colección de datos procedentes de estudios experimentales.
- Propone modelos matemáticos que pueden explicar algunos de los fenómenos biológicos estudiados.
- Proporciona soluciones matemáticas que permiten obtener predicciones para los procesos biológicos.
- Realiza estudios de comprobación de la calidad de los modelos descritos por medio de la comparación entre las simulaciones numéricas y los datos experimentales.

I.2 Biología Sintética

La biotecnología tradicional se define como toda aplicación tecnológica que utilice sistemas biológicos y organismos vivos o sus derivados, para la creación o modificación de productos o procesos en usos específicos [Snoep et al. (2006)].Por otra parte, la Ingeniería Genética consiste en la manipulación de la composición genética mediante la introducción o eliminación de genes específicos a través de técnicas de biología molecular y ácido desoxirribonucleico (ADN) recombinante. En otros términos, la Ingeniería Genética permite la edición del mensaje genético, es decir, la introducción de nuevas palabras que pueden cambiar el sentido de la frase. Cabe destacar entre sus primeros logros la producción de la insulina o la hormona de crecimiento humana a partir de cepas de la bacteria E.coli recombinante, sistemas que sustituirían a las fuentes alternativas que suponían su extracción a partir de cadáveres o la utilización de proteínas homólogas de otras especies animales.

El continuo avance de la Ingeniería Genética ha supuesto el nacimiento de un nuevo campo denominado Biología Sintética (BS). Esta ciencia se define como la síntesis de biomoléculas o ingeniería de sistemas biológicos con funciones nuevas que no se encuentran en la naturaleza [Heinemann and Panke, (2006)].Se trata de una nueva disciplina que, a diferencia de otras, no se basa en el estudio de la biología de los seres vivos, sino que posee un objetivo claro, el diseño de sistemas biológicos que no existen en la naturaleza, lo que hace que sea una disciplina que se sitúe más cerca de otras relacionadas con la ingeniería.

La estrategia de la BS consiste en emplear el conocimiento de los sistemas biológicos para diseñar nuevos sistemas biológicos con propiedades mejoradas o no existentes en la naturaleza. Esta estrategia es similar a la que permitió en su momento la expansión de la química orgánica, como nueva herramienta para la síntesis de nuevos compuestos no presentes en la naturaleza con propiedades de interés. La BS necesita un marco teórico que sea capaz de interpretar y predecir el comportamiento de los sistemas biológicos, lo que se consigue a través de Biología de Sistemas.

Por otra parte, una de las características principales de la Biología Sintética es su carácter interdisciplinar. Dentro de este campo emergente tienen cabida áreas de investigación y tecnologías asentadas, como la síntesis y secuenciación de ADN o la bioinformática, siempre y cuando sean aplicadas desde el enfoque sistemático y racional propio de la Biología Sintética.

Entre las áreas de interés que abarca esta nueva rama de la ciencia se pueden encontrar las siguientes:

- Biomedicina: se podrán obtener microorganismos capaces de fabricar complejos fármacos cuya fabricación actual se basa en fuentes naturales muy limitadas o costosos procesos de síntesis química. Por ejemplo la reconstrucción en levaduras, de un circuito genético encargado de la síntesis del precursor de la artemisina, un fármaco contra la malaria [Ro et al. (2006)]. Además, la BS permitirá el desarrollo de fármacos inteligentes que solo se activarán cuando se produzcan las circunstancias fisiológicas que requieran su participación, o posibilitará el desarrollo de fármacos personalizados capaces de segregar la cantidad necesaria de insulina en enfermos diabéticos de acuerdo a sus necesidades [Weiss, (2007)].
- Biorremediación: las aplicaciones medioambientales también salen beneficiadas de estos estudios. En el campo de la biorremediación los investigadores estudian la capacidad de los microorganismos evaluando su fisiología y utilizando una combinación del genoma usando técnicas experimentales y de modelización [Lovley, (2003)]. Otro ejemplo, en este sentido, lo constituye el diseño de biosensores, dispositivos de análisis capaces de reconocer e interaccionar con determinadas sustancias o microorganismos (como el sensor de arsénico [Aleksic et al. (2007)] o el sensor de TNT [Loogeret al. (2003)]).
- Biocombustibles: últimamente, otra área de investigación que está suscitando gran interés está relacionada con el campo de la producción de biocombustibles a partir de microorganismos fotosintéticos que sean capaces de utilizar un sustrato o directamente la luz solar para la producción de compuestos que puedan ser sustitutos del petróleo; en esta área, podemos citar la producción de hidrógeno [Navarro et al. (2009)], [Pinto et al. (2011)], [Triana et al. (2010)] o etanol [Pedrola et al. (En preparación)], [Hamelinck et al. (2005)]. También está la evolución hacia otros biocombustibles de más alta eficiencia y compatibilidad con la maquinaria actual como el butanol o incluso la fabricación de nuevos compuestos químicos más similares a los carburantes actuales [Montagud et al. (En preparación)]. Otras aristas de la producción de hidrógeno se puede encontrar en [Hallenbeck, (2002)].

Por tanto se puede afirmar que la BS se ha convertido en un área emergente de alto potencial científico y tecnológico e importancia estratégica tanto para la Unión

Europea, como en el más amplio contexto del desarrollo científico, tecnológico e industrial en los prolegómenos del siglo XXI. Correlativamente y en función precisamente de su carácter emergente, la Biología Sintética precisa de un gran impulso conceptual y metodológico para alcanzar su madurez como campo de desarrollo científico y tecnológico.

I.3 Bioinformática

La bioinformática es la rama de la ciencia en la cual la biología, las ciencias de la computación y las tecnologías de la información se mezclan para formar una sola disciplina. La meta de esta disciplina científica es permitir el descubrimiento de nuevas ideas, así como ofrecer una perspectiva global a partir de la cual se puedan discernir nuevos principios y paradigmas [Pevsner, (2009)]. En los comienzos de la revolución genómica, la bioinformática se restringía prácticamente a la creación y mantenimiento de bases de datos para almacenar información biológica, generalmente en la forma de secuencias nucleotídicas y aminoacídicas. El desarrollo de este tipo de bases de datos implicaba también el diseño de interfaces complejas que permitieran no solo el acceso a la información, sino la adición de nuevos datos y la revisión y actualización de los ya existentes. El desarrollo de nuevas vías para el manejo y gestión de la información biológica sigue siendo un objetivo fundamental de la bioinformática; sin embargo, el objetivo prioritario actual es el desarrollo de herramientas computacionales que permitan el análisis e interpretación de este gran volumen de información [Attwood et al. (2011)].

I.4 Técnicas computacionales

Las técnicas computacionales empleadas en Biología de Sistemas se basan en el desarrollo de algoritmos mediante los cuales la información obtenida por las técnicas experimentales es procesada y transformada en conocimiento [López et al. (2007)]. Gracias a estos algoritmos computacionales es posible realizar operaciones de *clustering* o agrupaciones de datos que poseen información semejante, o inferir la estructura o función de una proteína desconocida a partir de la secuencia e información estructural de proteínas homólogas, utilizando para ello un algoritmo de homología, por citar algún ejemplo. El objetivo final es realizar simulaciones de procesos biológicos o realizar modelos matemáticos, que permitan predecir el comportamiento de sistemas biológicos complejos y en último término, modificar o incluso diseñar estos sistemas en base a determinadas necesidades.

Las aproximaciones experimentales de proteómica, transcriptómica, genómica, etc., proporcionan gran cantidad de información. Los datos procedentes tan solo de una de estas aproximaciones proporcionan información acerca de los genes o de las proteínas únicamente, por lo que es necesaria una integración de estos datos por medio de una aproximación más global. Las técnicas computacionales proporcionan una herramienta que permite el tratamiento de toda esta información de manera integrada y permiten la mejora en las anotaciones acerca de las funciones de los genes y proteínas y además hacen posible la formulación de hipótesis biológicas.

I.5 Estructura de la tesis

La presente memoria para optar al grado de doctor se encuadra en la línea de investigación de Biología de Sistemas dentro del grupo de Modelización Interdisciplinar InterTech (www.intertech.upv.es). Está estructurada de la siguiente manera:

- En el Capítulo 1 se abordan los conceptos fundamentales relacionados con los modelos metabólicos. Estos representarían parcialmente un organismo virtual, mediante el cual se puede explorar posibles distribuciones de flujos dentro de la célula en diferentes condiciones medioambientales. Se exponen diversas herramientas y algoritmos para el análisis de los modelos metabólicos y se presentan algunos proyectos enfocados hacia la reconstrucción de modelos metabólicos de organismos con diferentes fines. Por último, se explican los pasos dentro de la metodología utilizada para la automatización del proceso de reconstrucción de modelos metabólicos a escala genómica. Se exponen en detalle los criterios probabilísticos de unicidad y completitud de las rutas metabólicas.
- En el Capítulo 2 se describen los conceptos de bases de datos y modelo relacional. Se analizan las nuevas tecnologías para la extracción de la información biológica disponible en diversas bases de datos biológicas. A continuación, se precisa la forma de obtener la información a partir de la construcción de un Servicio Web Cliente (SWC) que accede a la base de datos KEGG (del inglés, *Kyoto Encyclopedia of Genes and Genomes*). Tras hacer una valoración de los lenguajes y las herramientas utilizadas en la construcción del SWC, se detallan los artefactos construidos durante la ingeniería de software del mismo. Finalmente, se hacen algunas valoraciones sobre deficiencias encontradas en la información biológica. En

este sentido, se muestran resultados obtenidos durante la integración con otras bases de datos biológicas para suplir estas deficiencias.

- En el Capítulo 3 se describen los lenguajes y herramientas utilizadas en el diseño e implementación de COPABI para la reconstrucción de modelos metabólicos a escala genómica. Se precisa cada uno de los artefactos construidos durante la ingeniería de software de la aplicación.
- En el Capítulo 4 se exponen los resultados de la validación de los modelos metabólicos generados siguiendo la metodología que se expone en el Capítulo 1. La comparación se hace entre los modelos generados por COPABI y los publicados en la literatura para un mismo organismo. Durante el análisis de los modelos se miden características generales de las redes, conectividad de los nodos y semejanzas de las mismas a través del cálculo de un criterio definido para comparar redes metabólicas.
- En el Capítulo 5 se presentan las conclusiones de la investigación.

Capítulo 1

Modelos metabólicos

Para comprender el funcionamiento de un sistema biológico se ha de comenzar por realizar la descripción de sus componentes y posteriormente se debe emprender un análisis de su comportamiento frente a distintos estímulos. Con este objetivo, la Biología de Sistemas utiliza generalmente tanto modelos biológicos como aproximaciones computacionales. Los sistemas modelo que se utilizan preferentemente suelen ser organismos unicelulares tales como bacterias y levaduras, ya que poseen una complejidad menor que los mamíferos y son más sencillos de manipular.

El gran número de proyectos genómicos en curso, así como su alcance, indican la necesidad de herramientas totalmente automatizadas para generar nuevos conocimientos una vez completada la secuenciación del genoma. Los análisis experimentales producen gran cantidad de datos de tipo biológico que han de ser transformados en conocimiento, lo que se consigue mediante el empleo de diferentes tecnologías computacionales. En el presente capítulo se abordará la importancia de reconstruir modelos metabólicos a escala genómica de manera automatizada, así como las herramientas desarrolladas para este fin, haciendo una comparación con la metodología propuesta. Se describen los pasos implementados para la reconstrucción automática aplicando criterios probabilísticos.

1.1 Ingeniería metabólica

La ingeniería de vías metabólicas puede definirse como la modificación y/o introducción de reacciones bioquímicas, ya existentes o nuevas, para la mejora directa de propiedades celulares mediante tecnologías de ADN recombinante. Entre los propósitos que se persiguen se encuentran: mejorar rendimientos, extender el intervalo de sustratos metabolizables, extender el espectro de productos por complementación de vías, dirigir o reducir el flujo hacia una ramificación deseada, o bien mejorar productividades; esto se logra por medio de la re-estructuración de redes

metabólicas, la redistribución de flujos, o bien la amplificación, la eliminación o desregulación de genes [López et al. (2007)].

Algunas definiciones de ingeniería metabólica [López et al. (2007)]:

- Manipulación de los procesos metabólicos mediante tecnología recombinante con el objetivo de mejorar las propiedades de los microorganismos.
- Mejora dirigida de las propiedades celulares mediante ingeniería genética para modificar o introducir nuevas reacciones bioquímicas específicas.
- Alteración racional y dirigida de las rutas metabólicas de un organismo para comprender mejor y utilizar las rutas celulares en transformaciones (conversiones), transducción de energía y ensamblaje supramolecular.

La manipulación directa de las vías metabólicas en bacterias y otros microorganismos ha permitido el desarrollo y la mejora de cepas con la capacidad de sintetizar compuestos de utilidad como son algunos aminoácidos, polisacáridos, vitaminas, alcoholes, ácidos orgánicos, etc. Estas cepas han constituido la base para el desarrollo de nuevas y mejores tecnologías biológicas. Mediante la aplicación de la ingeniería de vías metabólicas, un número importante de compuestos orgánicos de uso industrial y terapéutico, obtenidos actualmente a partir de derivados del petróleo, pueden ser producidos por microorganismos utilizando otros carbohidratos además de glucosa como materia prima.

Las herramientas bioinformáticas así como las técnicas de ingeniería metabólica pueden emplearse para reconstruir las redes metabólicas de un microorganismo a partir del genoma únicamente [Deckwer, (2006)]. El objetivo no será solo el desarrollo de estrategias de cultivo en biorreactores a gran escala, sino además la determinación de las relaciones cinéticas de dependencia entre el genoma y el medioambiente con el fin de elaborar simulaciones de su comportamiento.

Por lo tanto, podemos afirmar que la Biología Sintética, como nueva disciplina en los límites entre la biología y la ingeniería, pretende el diseño parcial de organismos con fines aplicados, en base precisamente a la información disponible y a los métodos racionales de diseño en ingeniería, lo que demuestra la necesidad del establecimiento de un marco computacional y conceptual que proporcione asistencia en el desarrollo de sistemas biológicos artificiales, para lo que se hace necesario el desarrollo de nuevas herramientas computacionales integradas en un entorno común para el análisis de fenotipos metabólicos y el diseño de nuevos circuitos genéticos

complejos[Pacheco et al. (2011)]. En este sentido, la reconstrucción automática de modelos metabólicos a escala genómica es un objetivo primordial.

1.2 Modelos metabólicos

Uno de los objetivos fundamentales en el análisis en Biología de Sistemas es la reconstrucción de redes metabólicas a escala genómica. 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, con el objetivo de organizar la lista de reacciones metabólicas específicas para un organismo [Förster et al. (2003)]. En los últimos años los científicos han puesto a prueba diversos métodos y herramientas de gran utilidad en el desarrollo de dichas investigaciones. La variedad de aplicaciones de los modelos metabólicos incluye además, la búsqueda de sitios potenciales para la ingeniería metabólica, la determinación de las capacidades metabólicas, el diseño de estrategias óptimas de crecimiento, de producción de metabolitos, etc. [Oberhardt et al. (2009)].Si un modelo se formula adecuadamente, es de esperar que permita la simulación de los cambios en el metabolismo del organismo a través de perturbaciones ambientales y genéticas. Así, junto con las restricciones apropiadas, un modelo metabólico representaría parcialmente un organismo virtual, mediante el cual, y a través de análisis computacionales, se pueden explorar posibles distribuciones de flujos dentro de la célula en diferentes condiciones medioambientales y/o para una determinada configuración genética [Montagud et al.(2010)].

La reconstrucción de modelos metabólicos a escala genómica de organismos permite la integración de información genómica con actividades metabólicas observadas a través de experimentos fenotípicos y otras mediciones "ómicas" para obtener conocimiento biológico oculto y que pudiera ser de otro modo difícil de obtener [Fang et al. 2011)].

El proceso consiste en reunir toda la información referente al metaboloma de una especie, así como los genes que codifican a las enzimas que catalizan cada una de las reacciones metabólicas. Otros aspectos que se tienen en cuenta son las coenzimas y cofactores necesarios para la catálisis enzimática, la estequiometría y reversibilidad de las reacciones, información de la composición de la biomasa y aspectos de la regulación metabólica [Förster et al. (2003)].

En este contexto, se han desarrollado diversos proyectos enfocados hacia la reconstrucción de modelos metabólicos de organismos con diferentes fines, como la

producción optimizada de biocombustibles de tercera y cuarta generación mediante microorganismos como cianobacterias y levaduras [Montagud et al. (2010)], la reconstrucción del modelo metabólico de *Burkholderia cenocepacia J2315* para el estudio de tratamientos orientados a pacientes que padecen de Fibrosis Quística (FQ) [Fang et al. (2011)], la capacidad de *Rhodobacter sphaeroides* para producir hidrógeno, polihidroxibutirato u otros hidrocarburos, que lo convierten en un excelente candidato para su uso en una amplia variedad de aplicaciones biotecnológicas[Imam et al.(2011)], el estudio de *Clostridium beijerinckii* como un organismo que posibilita mejorar la producción de butanol, ya que (i) produce naturalmente las más altas concentraciones de butanol como subproducto de la fermentación, y (ii) puede lograr la co-fermentación de pentosayhexosa [Milne et al.(2011)]etc., en todos los casos demostrando la existencia de importantes lagunas, ya sea relacionado con las herramientas de cómputo disponibles para este fin así como por la incapacidad de las mismas para tener en cuenta criterios probabilísticos ofrecidos por el biólogo y que son muy útiles para lograr un modelo con calidad.

Con vistas a realizar análisis de los modelos metabólicos, están disponibles una variedad de herramientas y algoritmos [Patil et al. (2004)], incluyendo el análisis del balance de flujo (FBA, del inglés, *Flux Balance Analysis*) [Edwards et al. (1999)], [Jaime et al. (2010)], [Varma and Palsson, (1993)], minimización de los ajustes metabólicos (MOMA, del inglés *Minimization of Metabolic Adjustments*) [Segre et al. (2002)], el análisis del flujo metabólico (MFA, del inglés *Metabolic Flux Analysis*) [Schilling et al. (1999)], [Varma and Palsson, (1994)], así como herramientas para la visualización de redes metabólicas como el *Rational Organism Network Painter* [Garrido et al. (2010)].

Son varios los paquetes informáticos que asisten en la automatización de la reconstrucción de los modelos metabólicos a escala genómica. Múltiples esfuerzos muestran el interés de reducir a gran escala el tiempo requerido para llevar a cabo este proceso. Es preciso destacar algunos ejemplos como el desarrollado por Peter Karp, publicado como Pathways Tools en varias versiones. Esta aplicación consta de varios módulos: PathoLogic, Pathway Hole Filler, Pathway Tools Navigator y el Editor Functions. En esencia, comprende la utilización del genoma anotado de cualquier sistema biológico para inferir vías metabólicas probables existentes en el mismo, así como para generar una nueva base de datos de vías metabólicas y genomas [Karp et al. (2002)]. Recientemente, este autor y colaboradores, han expuesto un método para el rellenado múltiple de huecos en la red metabólica acelerando la generación de

modelos estequiométricos. Llamada *MetaFlux*, la aplicación se basa en la utilización de la programación MILP (del inglés, *Mixed Integer Linear Programming*) para corregir el conjunto de reacciones, metabolitos relacionados con la ecuación de biomasa (BM), nutrientes y procesos de secreción celular. El método genera los modelos metabólicos directamente de la base de datos de vías metabólicas y genomas gestionada por el Pathways Tools [Latendresse et al. (2012)].

Otros algoritmos son los implementados en las herramientas IdentiCS (del inglés, *Identification of Coding Sequences from Unfinished Genome Sequences*), metaSHARK (del inglés, *metabolic SearcH And Reconstruction Kit*) así como AUTOGRAPH (del inglés, *AUtomatic Transfer by Orthology of Gene Reaction Associations for Pathway Heuristics*). La aplicación IdentiCS consiste en el uso de secuencias genómicas no-anotadas para predecir regiones codificadoras y funciones asociadas a ellas, así como para la construcción *in silico* de la red metabólica [Sun and Zeng, (2004)]. El paquete informático metaSHARK se basa en la detección de genes codificadores de enzimas dentro del genoma no anotado de un sistema biológico y su visualización en el contexto de una red metabólica. Los dos módulos que lo componen SHARKhunt y SHARKview, ofrecen un nivel mejorado de flexibilidad y precisión en la automatización en la anotación de enzimas. El método implementado demuestra su utilidad en la generación de conocimientos en el metabolismo celular y por tanto en la reconstrucción de modelos metabólicos, a partir de genomas no anotados [Pinney et al. (2005)]. El método publicado como AUTOGRAPH se auxilia de la disponibilidad de los modelos metabólicos bien curados y publicados en artículos científicos para predecir de forma eficiente una equivalencia genética entre especies. Esto permite la transferencia hacia el modelo metabólico del organismo en estudio de la asociación gen/reacción a partir de la red metabólica publicada [Notebaart et al. (2006)].

Otras aplicaciones en este campo son las basadas en tecnología web, algunas de ellas son ReMatch, Model SEED y FAME. El algoritmo propuesto en ReMatch hace coincidir modelos metabólicos desarrollados por usuarios con la información de una base de datos interna, que incluye un glosario de nombres de metabolitos, generada a partir de KEGG, MetaCyc y CheBI. Por otro lado, ReMatch es capaz de aumentar el número de reacciones en el modelo, incorporadas a partir de la base de datos interna o introducidas por el usuario [Pitkänen et al. (2008)]. El recurso vía web Model SEED pretende analizar, comparar, reconstruir y curar la red metabólica a escala de sistemas biológicos. En este particular, los usuarios pueden enviar las secuencias genómicas en el sistema de anotación RAST y como resultado se construye una red de reacciones

metabólicas, establece una relación de asociación gen-proteína-reacción así como la BM para cada genoma analizado [Henry et al. (2010)]. Por otro lado, FAME es una herramienta de modelado que asiste en la creación, edición y análisis de modelos metabólicos de microorganismos a partir de la información de KEGG [Boele et al. (2012)].

Los servidores web constituyen otro recurso en la automatización de la reconstrucción de redes metabólicas. Cabe destacar algunos como MrBac (del inglés,*Metabolic network Reconstructions for Bacteria*), que intentan construir un borrador de la red metabólica. Esta herramienta integra análisis de genómica comparativa, recuperación de anotaciones del genoma así como la generación de archivos con formatos estándares en Biología de Sistemas [Liao et al. (2011)].

Las plataformas computacionales también ofrecen asistencia en esta labor. Ejemplos de estas son: MicrobesFlux y GEMSiRV (del inglés, *GEnome-scale Metabolic model Simulation, Reconstruction and Visualization*). MicrobesFlux es una plataforma web capaz de descargar automáticamente la red metabólica de aproximadamente 1200 especies depositadas en KEGG y convertirlas en borradores de los modelos metabólicos correspondientes. Además, la plataforma también proporciona herramientas personalizadas que permiten delecionar genes e introducir vías metabólicas heterólogas [Feng et al. (2012)]. Por su parte, GEMSiRV proporciona funcionalidades para la construcción y edición de redes metabólicas, para la visualización de redes integrando datos experimentales así como herramientas de análisis, por ejemplo FBA. Adicionalmente, todos los modelos metabólicos GEMSiRV-generados y resultados obtenidos, incluyendo proyectos en progreso, pueden ser fácilmente intercambiados por la comunidad científica [Liao et al. (2012)]. Otro ejemplo lo constituye el trabajo realizado por el grupo de investigación InterTech en el desarrollo de HYDRA (del inglés, *HYrbid Draw and Routes Analysis*), una plataforma computacional orientada al diseño, análisis y visualización de redes metabólicas [Garrido et al. (2011)], [Reyes et al. (2011)].

La pluralidad de estos recursos informáticos, a través de sus variados algoritmos, constatan un amplio empeño en automatizar el proceso de reconstrucción. Sin embargo, todavía persisten limitaciones en estos trabajos por lo que los resultados finales aún requieren los procesos de refinamiento manual. Si bien muchos de estos métodos gestionan de forma precisa la información en las grandes bases de datos, se siguen generando listados de reacciones no asociadas a las vías metabólicas con las que se relaciona. Adicionalmente, son incluidas reacciones metabólicas asociadas a

procesos genéticos moleculares, como es el caso de las reacciones de modificación y reparación del ADN, o con los procesos de replicación y transcripción del ADN o los procesos de división celular, reacciones de transferencia de señales, etc. Todas estas reacciones forman parte del metabolismo celular pero no son útiles a la hora de construir los modelos metabólicos, por lo que no deben incluirse.

De acuerdo con varios protocolos descritos en la literatura se requieren de 3 a 4 pasos fundamentales para la reconstrucción a escala genómica de modelos metabólicos [Thiele and Palsson, (2010)], [Triana et al., 2012)]. Este proceso pasa por generar un borrador del modelo metabólico del organismo en cuestión, seguido de un refinamiento exhaustivo (desde el principio o partir de ese paso se debe tener en cuenta el formato computacional para la escritura del modelo), así como de una validación del mismo a través de las capacidades de la red metabólica. La curación iterativa del modelo que incluye el análisis de huecos en la red así como el ajuste metabólico para su llenado, la eliminación de ciclos fútiles, entre otras, son parte crucial en el proceso de depurado. Se han reportado varios métodos que inciden en el refinamiento de la red [Osterman and Overbeek, (2003)], [Kharchenko et al. (2004)], [Kharchenko et al. (2006)], [Chen and Vitkup, (2006)], [Green and Karp, (2004)], [Kumar et al. (2007)]. Sin embargo, no existen algoritmos que definan criterios de completitud y unicidad para dichos modelos. Esta razón hace aún más engorroso el proceso de depuración si se tiene en cuenta que conlleva realizar un estudio de la reversibilidad de las reacciones, análisis de reacciones repetidas dentro del modelo, 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 su presencia también para los organismos de estudio.

Teniendo en cuenta esto, y a partir del estudio del primer modelo metabólico de un microorganismo fotosintético, la *Synechocystis sp. PCC6803* [Montagud et al. (2010)], obtenida por el grupo de Modelización Interdisciplinar InterTech, de la Universidad Politécnica de Valencia, pudimos comprobar la necesidad de desarrollar algoritmos que incluyan decisiones basadas en criterios probabilísticos (en todos los casos proporcionados por el biólogo) y que permitieran la reconstrucción de modelos metabólicos a escala genómica, cumpliendo estándares de coherencia y calidad de los modelos.

Estos criterios están basados en el análisis de la completitud y unicidad de las vías metabólicas. Para ello, en el primer caso requerirá la inclusión de vías metabólicas adicionales y su selección se basará en criterios probabilísticos (prevalencia de

metabolitos en un *pathway* general). Así mismo, la presencia múltiple de una misma reacción metabólica pero vinculada a diferentes enzimas debe ser depurada y seleccionar una única reacción. De nuevo, se usará un criterio probabilístico para la toma de decisión. Estos pasos serán explicados con más detalle en el próximo apartado.

1.3 Metodología para la reconstrucción del modelo metabólico:

Para la automatización del proceso de reconstrucción de modelos metabólicos a escala genómica en COPABI, se tuvo en cuenta la metodología implementada para la reconstrucción manual del modelo de la *Synechocystis sp. PCC6803* [Montagud et al. (2010)], donde además se incorporan los criterios probabilísticos tenidos en cuenta por los autores.

La siguiente figura refleja los pasos de la metodología implementada y a través de los cuales se obtiene un modelo metabólico a escala genómica depurado según los criterios probabilísticos que se tuvieron en cuenta:

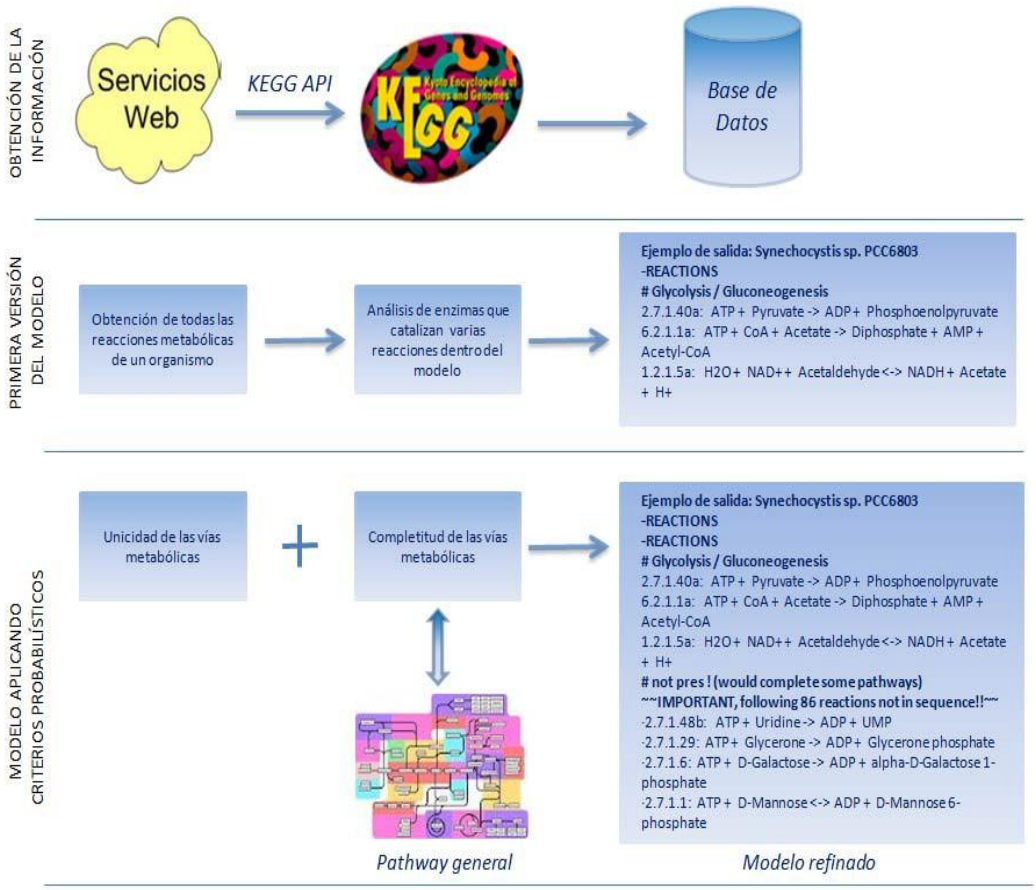

Figura 2. Diagrama general de la metodología implementada.

A continuación se explicarán los pasos en la automatización de dicha metodología:

1. El primer paso consistió en la compilación de todas las reacciones químicas de una ruta en particular o red metabólica en general de un organismo dado. En este caso se señala primeramente el nombre del *pathway* y a continuación el conjunto de reacciones metabólicas asociadas a él. Aquí se deben tener en cuenta varios elementos a la hora de organizar la información de las reacciones metabólicas:

    a) La mayoría de las reacciones están catalizadas por una enzima que posee un código llamado EC (del inglés, *Enzyme Commission*) (Ej: 1.1.1.1). Este código se coloca al principio de cada reacción. (Cabe señalar que asumimos poner el número EC al principio de cada reacción por un problema de conveniencia particular; se puede poner otro identificador).
    b) Hay reacciones en varios pathways que no están catalizadas por enzimas, o sea que son reacciones espontáneas. Esas reacciones aparecen en la base de datos con el nombre: "*non-enzymatic*" y de la misma manera debe aparecer en el modelo que se genere. O sea, se coloca "*non-enzymatic*" al principio de la reacción para especificar que estamos en presencia de una reacción espontánea.

    En este paso es primordial tener en cuenta la reversibilidad de la reacción, pues esto puede afectar los resultados finales en el análisis del modelo.

    Ejemplo de salida: Synechocystis sp. PCC6803

    -REACTIONS

    # Glycolysis / Gluconeogenesis

    1.2.4.1a: Pyruvate + Thiamin diphosphate -> CO2 + 2-(alpha-Hydroxyethyl) thiamine diphosphate

    2.7.1.40a: ATP + Pyruvate -> ADP + Phosphoenolpyruvate

    6.2.1.1a: ATP + CoA + Acetate -> Diphosphate + AMP + Acetyl-CoA

    1.2.1.5a: H2O + NAD+ + Acetaldehyde <-> NADH + Acetate + H+

    1.1.1.2: NADP+ + Ethanol <-> NADPH + H+ + Acetaldehyde

    # Citrate cycle (TCA cycle)

    1.1.1.42a: Oxalosuccinate <-> CO2 + 2-Oxoglutarate

    1.1.1.37: NAD+ + (S)-Malate -> NADH + Oxaloacetate + H+

2.3.3.1: CoA + Citrate <-> H2O + Acetyl-CoA + Oxaloacetate
6.2.1.5a: ATP + CoA + Succinate <-> ADP + Orthophosphate + Succinyl-CoA

2. El segundo paso está relacionado con la identificación del conjunto de enzimas, donde cada una de ellas puede catalizar varias reacciones del mismo tipo pues sus sustratos solo presentan pequeñas diferencias en sus estructuras químicas. En ese caso se pone el identificador de la enzima y una letra para denotar que esa enzima cataliza varias reacciones o un número cuando es una reacción espontánea (por ejemplo: 1.1.1.1a, 1.1.1.1b, non-enzymatic1, non-enzymatic2, etc.).

Hasta aquí hemos obtenido exactamente la información biológica que está almacenada en la base de datos. En lo adelante será necesario introducir de manera automática los criterios probabilísticos para satisfacer los criterios de unicidad y completitud exigibles al mapa metabólico.

Criterios probabilísticos a tener en cuenta para reconstruir el modelo metabólico:

Para el desarrollo del algoritmo capaz de reconstruir el modelo metabólico a escala genómica de cualquier organismo se tuvieron en cuenta criterios probabilísticos que garantizaran la completitud y unicidad de las vías metabólicas. Esto implica establecer criterios para la eliminación de una reacción metabólica y evitar su repetición dentro del modelo y la incorporación de nuevas reacciones a partir de la comparación con un *pathway* general generado a partir de la compilación de todas las reacciones metabólicas que existen en los organismos en la naturaleza.

Análisis de cada criterio:

a. Unicidad de las vías metabólicas:

La eliminación de una reacción metabólica para evitar su repetición dentro del modelo implica analizar la existencia de una misma reacción en varias rutas metabólicas dentro del modelo metabólico, pues las reacciones deben aparecer solo una vez en el mismo.

En este caso, se propone calcular la cantidad de veces que se repite la enzima que cataliza cada una de ellas en todo el modelo del organismo y, una vez

obtenidos estos valores, se comparan y se mantiene en el modelo la reacción que está relacionada con la enzima que menos veces aparezca en todo el modelo. Ejemplo:

1.2.1.12 a: A + B <-> C + D

1.2.1.12 b: G + E <-> K + L

1.2.1.12 c: P + V -> Y

1.1.1.1: G + E <-> K + L

Como se puede apreciar, las reacciones subrayadas son iguales, de ahí que se calcule la cantidad de veces que aparece en todo el modelo las enzimas asociadas a ambas reacciones (1.2.1.12 = 3) y (1.1.1.1 = 1). Luego se comparan ambos valores, llegando a la conclusión de que la enzima (1.2.1.12) ya está incluida en el modelo y la enzima 1.1.1.1 se encontró que está en la anotación del organismo y la única reacción que cataliza es la (G + E <-> K + L).

Por tanto se mantiene la reacción G + E <-> K + L asociada a la enzima (1.1.1.1).

b. Completitud de las vías metabólicas

La incorporación de nuevas reacciones en el modelo metabólico está asociada a la comparación con un *pathway* general (teórico) que fue generado a partir de la compilación de todas las reacciones metabólicas que existen en los organismos en la naturaleza. En este caso, los biólogos advierten sobre la existencia de genes, y por tanto de reacciones, que no están en las bases de datos biológicas, es decir, que no se han reportado en la literatura, y que se pueden adicionar en los modelos.

Por ejemplo: un *pathway* X, que consta, por estudios bioquímicos, de 10 reacciones teóricas, y se tiene que para un organismo Y están presentes en la base de datos solo 8 reacciones de las 10 posibles. En este caso, se ordenan secuencialmente cada una de las 8 reacciones y se incorporan al modelo. El resto de las reacciones que no aparecían descritas para ese organismo se agregan en el modelo (y se hace la salvedad de que no están reportadas en la literatura, o sea en la base de datos) cuando cumplen en conjunto las siguientes condiciones:

1) Son reacciones cuyo producto final es un metabolito que se computa dentro de la BM.

La interfaz de COPABI permite al usuario seleccionar dentro de un listado de metabolitos aquellos que formarán parte de la BM, además de poder añadir otros. En todos los casos deberá introducir el coeficiente estequiométrico asociado a cada metabolito y que servirá para luego conformar la BM.

2) dicho metabolito debe aparecer repetido con una frecuencia dada en cierto porcentaje dentro del *pathway* general.

La interfaz de COPABI permite al usuario introducir un parámetro de decisión dado en porcentaje, que permitirá valorar la presencia de los metabolitos especificados en el punto 1 dentro del *pathway* general. Finalmente, el algoritmo implementado permite adicionar al modelo que se construye todas las reacciones dentro del *pathway* general que cumplen con las condiciones anteriores de manera simultánea y con el parámetro de decisión definido por el usuario.

La forma de identificar en el modelo generado las reacciones que fueron incorporadas es: **# not pres ! ~~IMPORTANT, following X reactions not in sequence!!~~** . Además a cada reacción le antecede un punto (.) que significa que son reacciones agregadas en el modelo para lograr la completitud de las vías metabólicas.

Ejemplo de salida: Synechocystis sp. PCC6803

 # not pres ! (would complete some pathways) ~~IMPORTANT, following 10 reactions not in sequence!!~~

·6.3.2.2: ATP + L-Glutamate + L-Cysteine -> ADP + Orthophosphate + gamma-L-Glutamyl-L-cysteine

·6.3.4.3: ATP + Formate + Tetrahydrofolate <-> ADP + Orthophosphate + 10-Formyltetrahydrofolate

·2.7.1.48b: ATP + Uridine -> ADP + UMP

·2.7.1.29: ATP + Glycerone -> ADP + Glycerone phosphate

·2.7.1.6: ATP + D-Galactose -> ADP + alpha-D-Galactose 1-phosphate

·2.7.1.1: ATP + D-Mannose <-> ADP + D-Mannose 6-phosphate

·2.7.1.16a: ATP + D-Ribulose <-> ADP + D-Ribulose 5-phosphate

·2.7.1.45: ATP + 2-Dehydro-3-deoxy-D-gluconate -> ADP + 2-Dehydro-3-deoxy-6-phospho-D-gluconate

·2.7.1.48c: dATP + Cytidine -> CMP + dADP

·2.7.1.48d: dATP + Uridine -> UMP + dADP

Aspectos importantes:

El *pathway* general (teórico) es muy importante pues ahí aparecen todas las rutas que existen en los organismos en la naturaleza, es decir, es como la enciclopedia de las reacciones químicas en el Metabolismo Celular. Este pathway general constituye el patrón para comparar pues, al estar todas las reacciones, se puede tomar como referente para comparar cuántas reacciones le faltan a una ruta X en un organismo Y.

Por otra parte, es válido señalar que las anotaciones no son perfectas y que siempre quedan genes por identificar.

Otros elementos dentro del modelo:

Las vías metabólicas constituyen un paradigma central en la biología. Históricamente, se han definido sobre la base del descubrimiento de cada reacción constituyente. Sin embargo, las redes metabólicas a escalas genómicas, actualmente reconstruidas a partir de la anotación de secuencias del genoma de cada organismo, demandan nuevas definiciones basadas en redes para estas vías metabólicas con el objetivo de facilitar el análisis de sus capacidades y funciones, tales como: la versatilidad y robustez metabólica, las velocidades óptimas de crecimiento, etc. Esta demanda ha llevado a la aplicación de análisis matemáticos complejos al desarrollo de métodos que permiten predecir, modelar y simular comportamientos fenotípicos determinados. Hoy en día muchos de los métodos de análisis se han desarrollado bajo una aproximación basada en restricciones. Estos métodos basados en restricciones, que permiten el análisis de los estados fenotípicos de microorganismos a escala genómica han sido desarrollados rápidamente en los años recientes. En sentido general la implementación de estos métodos consiste en varios pasos; en primera instancia se lleva a cabo la reconstrucción de la red metabólica a escala genómica. Como segunda consideración, es imprescindible la aplicación de restricciones apropiadas para la construcción *in silico* del modelo a escala genómica. Por último, se utilizan varios algoritmos de análisis para evaluar las propiedades de estos modelos. Muchos de estos estudios están basados en la utilización de métodos de optimización como son la programación lineal, la cuadrática y la no lineal, los cuales implican la definición de un conjunto de restricciones para su ejecución. Por otro lado, la solución buscada a través de estos métodos de optimización está fijada en la forma de una

función objetivo a optimizar. La representación general de esta función objetivo permite la formulación de rangos de funcionalidades y estados de la red de interés. La misma puede usarse para representar la búsqueda de las capacidades metabólicas de la red, objetivos de relevancia fisiológica (como por ejemplo: el máximo de la velocidad de crecimiento celular o de la biomasa), o el diseño de objetivos para la producción de un metabolito de interés. La optimización de la formación de biomasa es uno de los objetivos más ampliamente utilizados para determinar la máxima velocidad de crecimiento celular bajo determinadas condiciones ambientales [Edwards et al. (2001)], [Ibarra et al. (2002)].

Los metabolitos externos, por su parte, son las fuentes y los sumideros de la red metabólica; son los que "alimentan" todo el sistema, por lo tanto, estos compuestos no son balanceados dentro del análisis que se realiza a las vías metabólicas ya que sus concentraciones por lo general varían con el tiempo. En varios análisis, los metabolitos internos, a diferencia de los externos, tienen que cumplir la condición de estado estacionario dentro del sistema [Dandekar et al.(2003)].

Dicho esto, y teniendo en cuenta la importancia de estos elementos para el análisis posterior de los modelos metabólicos a escala genómica, COPABI presenta en su interfaz la opción de poder conformar la BM, proporcionando un listado de los metabolitos que pudieran formar parte de la misma, ya sea actuando como reactante o como producto, y con el que el usuario conforma la BM correspondiente. Además da la posibilidad al usuario de poder incorporar nuevos metabolitos que no se encuentran inicialmente en la aplicación. En todos los casos tendrá que indicar el coeficiente estequiométrico de cada uno de ellos. Ejemplo de cómo se muestra en el modelo generado es el siguiente:

BM: -> 0.765L-alanine + 0.456L-aspartate + 0.3455L-arginine + 1.234Glucose -> 1.876Biomass + 2.456ADP + 0.234Orthophosphate + 1.876product

De la misma manera COPABI tiene en su interfaz la opción de que el usuario pueda incluir los metabolitos externos y las restricciones del modelo metabólico, que serán el punto de partida para posteriores análisis con los modelos generados.

Capítulo 2

Obtención de la información biológica

En los últimos años, las aplicaciones de la Biotecnología en diferentes ámbitos de la Ciencia y la Tecnología se han multiplicado de forma considerable en paralelo al incremento exponencial de la información que se posee sobre los organismos, entre otros sobre su genética, sus procesos de regulación y su metabolismo. Dicha información, obtenida mediante diferentes tecnologías cada vez más poderosas y eficientes, está pasando a formar parte de grandes bases de datos, muchas de ellas de libre acceso, que, en combinación con el vasto conjunto de publicaciones científicas, ponen en manos de los investigadores una gran cantidad de datos.

Un aspecto primordial para obtener los modelos metabólicos es la búsqueda de la información biológica necesaria, la cual se encuentra almacenada en bases de datos públicas, como por ejemplo: Biocyc [Karp et al. (2005)], KEGG [Kanehisa et al. (2008)], Brenda [Chang et al. (2009)] o Uniprot [Wu et al. (2006)], [The UniProt Consortium, (2007)] y que puede ser recopilada para un organismo específico. Sin embargo, la falta de calidad debe ser considerada como uno de los principales inconvenientes de algunas de las bases de datos: falsos positivos, falsos negativos, así como objetos anotados erróneamente pueden obstaculizar los esfuerzos para reunir datos exactos [Weise et al. (2006)]. En consecuencia, la reconstrucción a partir de un control minucioso de todas y cada una de las reacciones, la BM basada en moléculas constituyentes (como aminoácidos y nucleótidos) o la coherencia y la integridad de la red son requisitos previos para la generación de un modelo metabólico de alta calidad y útil [Feist et al.(2009)].

A continuación se analizarán los conceptos de bases de datos y modelos de datos, en aras de identificar la mejor variante en cada caso para almacenar la información biológica necesaria para reconstruir automáticamente los modelos metabólicos. Además se explicarán las tecnologías y herramientas utilizadas en la construcción de

un SWC para la conexión y descarga de dicha información a partir de la base de datos KEGG.

2.1 Bases de datos

Una base de datos es un conjunto de datos interrelacionados entre sí, o sea, una colección de datos variables en el tiempo. El software que permite la utilización y/o la actualización de los datos almacenados en una (o varias) base(s) de datos por uno o varios usuarios se denomina Sistema Gestor de Base de Datos (SGBD) [Elmasri and Navathe, (2007)].

Como resultado del desarrollo tecnológico de campos como la informática y la electrónica, en la actualidad la mayoría de las bases de datos están en formato digital, lo cual ofrece un amplio rango de soluciones al problema de almacenar datos.

Tipos de bases de datos según variabilidad de los datos:

- Bases de datos estáticas: son bases de datos de solo lectura, utilizadas primordialmente para almacenar datos históricos que posteriormente se pueden utilizar para estudiar el comportamiento de un conjunto de datos a través del tiempo, realizar proyecciones y tomar decisiones.
- Bases de datos dinámicas: son bases de datos donde la información almacenada se modifica con el tiempo, permitiendo operaciones como actualización y adición de datos, además de las operaciones fundamentales de consulta. Un ejemplo de esto puede ser la base de datos utilizada en una farmacia, un videoclub, etc.

2.1.1 Modelos de datos

Un modelo de datos es básicamente una "descripción" de algo conocido como contenedor de datos, así como de los métodos para almacenar y recuperar información de esos contenedores. Los modelos de datos no son elementos físicos: son abstracciones que permiten la implementación de un sistema eficiente de base de datos; por lo general se refieren a algoritmos, y conceptos matemáticos. En este sentido, no es solamente un modo de estructurar datos, sino que también define el conjunto de operaciones que pueden ser realizadas sobre los datos [Korth et al. (2006)]. A continuación se hace referencia al modelo de

datos utilizado en el diseño de la base de datos desarrollada para almacenar la información biológica necesaria para nuestro grupo de investigación.

2.1.2 Modelo Relacional

Una base de datos relacional es una base de datos que cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para modelar problemas reales y administrar datos dinámicamente. Permite establecer interconexiones (relaciones) entre los datos y trabajar con ellos conjuntamente. Tras ser postuladas sus bases en 1970 por Edgar Frank Codd, no tardó en consolidarse como un nuevo paradigma en los modelos de base de datos [Pons et al. (2005)].

El lenguaje más común para construir las consultas a bases de datos relacionales es SQL (del inglés, *Structured Query Language*), un estándar implementado por los principales motores o sistemas de gestión de bases de datos relacionales [Date and Darwen, (2009)].

2.1.3 Bases de Datos Biológicas

Una de las más visibles consecuencias del paso de la era genómica a la post-genómica fue el nacimiento de una comunidad de información biológica. Actualmente son las bases de datos relativas a la biología las que más rápido crecimiento tienen y en las que más tiempo de desarrollo se invierte [Reyes et al. (2010)]. Las bases de datos biológicas nacen como un intento de recopilar y permitir el libre acceso a la información por parte de la comunidad de investigadores, facilidades que se fueron haciendo poco a poco de uso “común” entre la comunidad de biólogos, hasta llegar a un concepto relativamente nuevo en la implementación de bases de datos a través de Internet, la integración de herramientas de comparación y análisis de secuencias con las mismas bases de datos.

Una base de datos biológica es una biblioteca de información sobre ciencias de la vida, recogida de experimentos científicos, literatura publicada, tecnología de experimentación de alto rendimiento y análisis computacional. Contiene información de áreas de investigación incluyendo genómica, proteómica, metabolómica, expresión génica y filogenética [Altman, (2004)].La información

contenida en bases de datos biológicas incluye funciones, estructura y localización (tanto celular como cromosómica) de genes, efectos clínicos de mutaciones, así como similitudes de secuencias y estructuras biológicas.

El diseño de estas bases de datos, su desarrollo y su gestión a largo plazo, forman un área nuclear de la disciplina de la bioinformática [Bourne, (2005)]. El contenido de los datos incluye secuencias génicas, descripciones textuales, atributos y clasificaciones ontológicas, anotaciones, y datos en forma tabular. En la actualidad están disponibles sistemas que permiten consultar simultáneamente múltiples bases de datos. Estos recursos no solo permiten acceder a las secuencias de interés y a la información básica de ellas, sino que en una sola búsqueda se puede recopilar información relacionada, probablemente disponible en otras bases de datos, como información taxonómica del organismo a partir del cual fueron extraídas las secuencias; características de la estructura tridimensional en el caso de las proteínas o información acerca de los genes específicos, como la posición en el genoma del organismo en cuestión o su posible asociación con patologías humanas.

Ejemplos de Bases de Datos Biológicas

Enzyme: es una base de datos relacionada con la nomenclatura de las enzimas. Está basada en las recomendaciones del IUBMB (del inglés, *Nomenclature Committee of the International Union of Biochemistry*) y describe cada tipo de enzima caracterizada [Bairoch, (2000)].

GenBank: es una colección pública de secuencias tanto de proteínas como de ácidos nucleicos con soporte bibliográfico (referencias tomadas de la literatura reportada) y notación biológica (especie y origen). La base de datos del GenBank crece de una manera exponencial; este crecimiento es debido a la forma en que la base se actualiza. Son los mismos autores quienes se encargan de mantener la base al día, pero además, el GenBank se nutre de las otras bases de datos existentes actualizando sus ficheros [Miller et al. (2009)].

MetaCyc: es una base de datos no redundante de rutas metabólicas dilucidadas de manera experimental. En estos momentos cuenta con más de 1100 rutas de más de 1500 organismos las cuales están involucradas tanto en metabolismo primario como secundario, compuestos asociados, enzimas y genes. Puede ser usada en una gran variedad de aplicaciones científicas como: proveer

datos de referencia para predicciones computacionales en rutas metabólicas de organismos con genomas secuenciados, soporte para ingeniería metabólica, ayuda a la comparación entre redes metabólicas y como una enciclopedia de metabolismos. Las rutas pueden ser buscadas mediante una lista, con ontologías, o haciendo una consulta de manera directa preguntando por las rutas, proteínas, reacciones o compuestos [Caspi et al. (2008)], [Krieger et al. (2004)].

HPRD (Protein Reference Database): toda la información depositada en HPRD ha sido extraída de manera manual de la literatura por biólogos expertos, los cuales leen, interpretan y analizan los datos publicados. La base de datos fue creada utilizando una base de datos orientada a objetos, lo cual proporciona versatilidad en su consulta y permite que los datos sean desplegados de manera dinámica [Peri et al. (2003)].

WikiPathways: es un sitio que facilita la contribución y el mantenimiento de información relacionada con las vías biológicas dentro de la comunidad biológica. Esta base de datos es un proyecto abierto y colaborativo que fue publicado en el 2008 con el propósito de brindar a la comunidad una plataforma para el depurado de vías biológicas. Este sitio tiene herramientas para editar estas mismas vías [Kelder et al. (2009)].

KEGG: es una base de datos de sistemas biológicos que está compuesta por elementos genéticos, fundamentalmente de genes y proteínas (*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*). *KEGG* provee de un conocimiento base para relacionar genomas tanto con sistemas biológicos como con el ambiente haciendo procesos de *mapeo* con *PATHWAY* y *BRITE* [Kanehisa et al. (2008)].

Brenda: es la colección principal de información sobre la función de enzimas disponible para la comunidad científica [Chang et al. (2009)].

2.2 Tecnologías para la construcción del Servicio Web Cliente

Hoy en día existen millones de páginas y grandes cantidades de información en el World Wide Web (WWW). Con la llegada de Internet, la demanda de sitios Web comenzó a crecer rápidamente y muchas organizaciones descubrieron el potencial de la Web, que prontamente se convirtió en un medio grande y poderoso para que las

empresas estuvieran en contacto con clientes y proveedores, expresaran opiniones y sacaran provecho de las aplicaciones de comercio electrónico. WWW fue creado originalmente por científicos para compartir información y documentos. Los sitios Web necesitan ser flexibles y proveer funcionalidades dinámicas para interactuar con otras aplicaciones existentes. El típico entorno de desarrollo Web necesita una combinación de diferentes tecnologías, herramientas y arquitecturas.

Las oportunidades que ofrece este crecimiento de Internet han motivado un intenso trabajo en el ámbito de la investigación y en el de la industria. Flexibilidad, interconectividad, autonomía e independencia de plataforma juegan un rol fundamental en el desarrollo de software. El software se está convirtiendo cada vez más en un servicio ofrecido a los usuarios o a otros elementos software, y no se ve como una aplicación aislada ejecutándose en una máquina específica para un requisito predefinido. Esta es la visión del software como "servicio" bien conceptualizado por el paradigma de Computación Orientada a Servicios (SOC, del inglés *Service Oriented Computing*).

Las aplicaciones más conocidas del paradigma de SOC se pueden encontraren la Web: la Arquitectura Orientada a Servicios (SOA, del inglés *Service Oriented Architecture*) y los Servicios Web (WS, del inglés *Web Service*). (Ver figura 3).

Figura 3. Relación entre SOC, SOA y los servicios Web.

2.2.1 Arquitectura Orientada a Servicios

Existen múltiples definiciones de SOA [Crawford et al. (2005)], pero la más aceptada es la proporcionada por [Bell, (2008)]: "SOA es un conjunto de componentes

que pueden ser invocados, cuyas descripciones de interfaces se pueden publicar y descubrir". Una arquitectura orientada a servicios tiene varios elementos fundamentales que deben aparecer. El principal concepto es “servicio” y las colaboraciones principales son “publicar”, “descubrir/buscar” e “interactuar”.

Roles de SOA

Existen tres roles dentro de SOA (ver figura 4):

- Proveedor de servicios: implementa el servicio y lo hace accesible en Internet.
- Consumidor de servicios: interactúa con un proveedor de servicios. Tradicionalmente se le llama “cliente”. Puede ser una aplicación final u otro servicio.
- Repositorio de servicios: provee un lugar donde los desarrolladores pueden publicar nuevos servicios y buscar otros existentes.

Figura 4. Interrelación entre los roles de SOA.

Cada entidad en SOA, puede jugar uno o más de los tres roles de proveedor, consumidor o repositorio de servicios. En la figura 4 se muestran los tres tipos de colaboración entre los roles:

- Publicar servicio: un proveedor de servicios publica un servicio, haciéndolo disponible a los consumidores a través de un repositorio de servicios.
- Buscar servicio: los consumidores de servicios buscan y localizan los servicios en un repositorio de servicios.

- Interactuar: es la comunicación entre el consumidor y los servicios del proveedor. El consumidor realiza peticiones al servicio a través de los protocolos que indica la información del servicio que tiene el repositorio.

2.2.2 Servicios Web

Una aplicación SOA está formada por un conjunto de servicios que encapsulan los procesos de negocio. Los servicios realizan funciones que pueden ir desde la más simple respuesta hasta el más complicado proceso de negocio. Permiten a las organizaciones exponer su funcionalidad sobre Internet usando lenguajes y protocolos estándares y son implementados mediante interfaces, basadas en estándares abiertos.

Las instancias más conocidas de servicios, son los servicios Web. Estos proporcionan la plataforma tecnológica ideal para conseguir la completa integración de los procesos de negocio de una organización con diferentes organizaciones. Los servicios Web prometen ser el mecanismo adecuado para la implementación de SOA en sistemas integrados y distribuidos. A continuación se presenta cómo se define un WS, sus características, arquitectura, protocolos y estándares que lo definen.

Definición

Existen múltiples definiciones sobre lo que son los servicios Web, lo que muestra su complejidad a la hora de dar una adecuada definición que englobe todo lo que son e implican. Sin embargo, una definición bastante completa es la siguiente:

“Un servicio web permite que distintas aplicaciones de software desarrolladas en lenguajes de programación diferentes, y ejecutadas sobre cualquier plataforma, puedan utilizar los servicios web para intercambiar datos en redes de ordenadores como Internet.”[Benslimane et al. (2008)].

Los servicios Web facilitan el acceso a la funcionalidad de las aplicaciones a través de internet, facilitando la interoperabilidad entre servicios y aplicaciones y permitiendo integrar la funcionalidad de distintas aplicaciones. Además, proporcionan estándares y mecanismos para llevar a cabo el comercio electrónico, convirtiendo la Web en un marco ideal para el desarrollo de aplicaciones distribuidas en prácticamente todos los dominios de aplicación [Snell et al. (2001)]. Los servicios Web constituyen el principal mecanismo para implementar la Arquitectura Orientada a Servicios.

2.2.3 XML

XML (del inglés, *Extensible Markup Language*) es un lenguaje extensible de etiquetas que se utiliza para normalizar el intercambio de datos entre participantes, proporcionando un medio para codificar y formatear los datos. Es similar a HTML (del inglés, *HyperText Markup Language*), con elementos, atributos y valores. Sus elementos y atributos definen tipos y estructuras de información para los datos que llevan, incluyendo la capacidad de modelar datos y estructuras específicas de un dominio de sistema [Harold, E.R. (2003)]. Un aspecto fundamental de los servicios Web es transformar un XML genérico de datos en una aplicación o en una representación de dominio específico de datos. La sintaxis de XML usada en las tecnologías de los WS especifica cómo se representan los datos, define cómo y con qué calidad se transmiten los datos y los detalles de cómo se publican y descubren los servicios.

2.2.4 SOAP

SOAP (del inglés, *Simple Object Access Protocol)* es un protocolo estándar que define la comunicación entre dos objetos a través del intercambio de datos. Proporciona un modo abierto y extensible para que las aplicaciones se comuniquen a través de la Web usando mensajes basados en XML, con independencia de sistemas operativos, modelos de objetos o lenguajes de programación [Snell et al. (2001)].

Facilita la comunicación universal, definiendo un formato de mensajes simple y extensible en XML estándar y proporcionando además, un modo para enviar esos mensajes sobre el protocolo de comunicación HTTP (del inglés, *Hypertext Transfer Protocol*). SOAP intercambia información mediante mensajes y estos se utilizan como envoltorios que la aplicación utiliza para guardar la información que quiere enviar.

2.2.5 WSDL

WSDL (del inglés, *Web Services Description Language)* es un lenguaje basado en XML que se utiliza para describir las funcionalidades de los servicios Web. Permite separar la descripción de la funcionalidad abstracta ofrecida por un servicio de los detalles concretos de la descripción del servicio. El documento WSDL de un servicio, proporciona dos piezas de información básicas: (1) una

parte o interfaz abstracta (independiente de la aplicación), y (2) una parte concreta que define los enlaces a protocolos e información de los puntos finales de acceso al servicio [Christensen et al. (2009)].

De esta forma, WSDL se utiliza para describir un WS en términos de los mensajes que acepta y genera, actúa como contrato entre un consumidor (cliente) y dicho servicio.

2.2.6 Servicios web en bioinformática

Se han desarrollado interfaces basadas en SOAP para una amplia variedad de aplicaciones bioinformáticas, permitiendo que una aplicación, corriendo en un ordenador de cualquier parte del mundo, pueda usar algoritmos, datos y recursos de computación alojados en distintos servidores. La principal ventaja radica en que el usuario final no tiene que ocuparse de actualizaciones y modificaciones en el software o en las bases de datos [Harte et al. (2004)]. Entre los servicios web disponibles se pueden encontrar los siguientes:

- Servicios de obtención de información en línea, como por ejemplo: consultas a bases de datos.
- Herramientas de análisis.
- Búsquedas de similitudes entre secuencias.
- Alineamientos múltiples de secuencias.
- Análisis estructural.
- Servicios de acceso a literatura especializada y ontologías.

2.2.7 KEGG API

KEGG API es un WS que permite usar el sistema de KEGG vía SOAP/WSDL. Proporciona valiosos medios para acceder a la información disponible en dicha base de datos, tanto para la búsqueda en los procesos bioquímicos celulares o analizar el universo de los genes en los genomas completamente secuenciados [Reyes et al. (2011)]. Los usuarios pueden acceder al servidor de KEGG API por la tecnología SOAP a través del protocolo HTTP. El servidor SOAP también viene con el WSDL, lo que hace que sea fácil de construir una biblioteca de cliente para un lenguaje de programación específico. Esto permite a los usuarios escribir sus propios programas para diversos propósitos y para automatizar el procedimiento

de acceso al servidor KEGG API y recuperar los resultados. Entre los principales métodos que brinda este servicio se encuentran los siguientes:

- **list_organisms**: devuelve el listado de los organismos presentes en la base de datos KEGG/GENES.
- **list_pathways**: devuelve el listado de *pathways* correspondientes a un organismo dado.
- **get_genes_by_organism:** devuelve todos los genes del organismo especificado.
- **get_number_of_genes_by_organism**: devuelve el número de genes codificados en el genoma del organismo especificado.
- **get_genes_by_pathway:** busca todos los genes del *pathway* especificado**.**
- **get_compounds_by_pathway:** busca todos los compuestos presentes en el *pathway* especificado**.**
- **get_reactions_by_pathway**: busca todas las reacciones presentes en el *pathway* especificado**.**
- **get_enzymes_by_pathway**: busca todas las enzimas presentes en el *pathway* especificado**.**
- **get_reactions_by_enzyme**: devuelve todas las reacciones que están vinculadas con una enzima dada.
- **bget:** permite obtener todos los datos correspondientes al elemento biológico deseado, devolviendo la información en un *array* de *string.*

Después de un análisis sobre la información biológica a extraer de KEGG y partiendo de los métodos que dicha base de datos propone a partir de su WS, fue posible la construcción de un SWC que se conecta a KEGG y extrae la información biológica necesaria para la reconstrucción automática de modelos metabólicos a escala genómica [Reyes et al. (2011)].

En la siguiente figura se muestra la secuencia de descarga de la información biológica y los métodos utilizados para obtener la información desde KEGG. En todos los casos fue necesario utilizar el método *bget* para obtener información adicional de las entidades. Por parámetro se pasan los valores específicos para cada función.

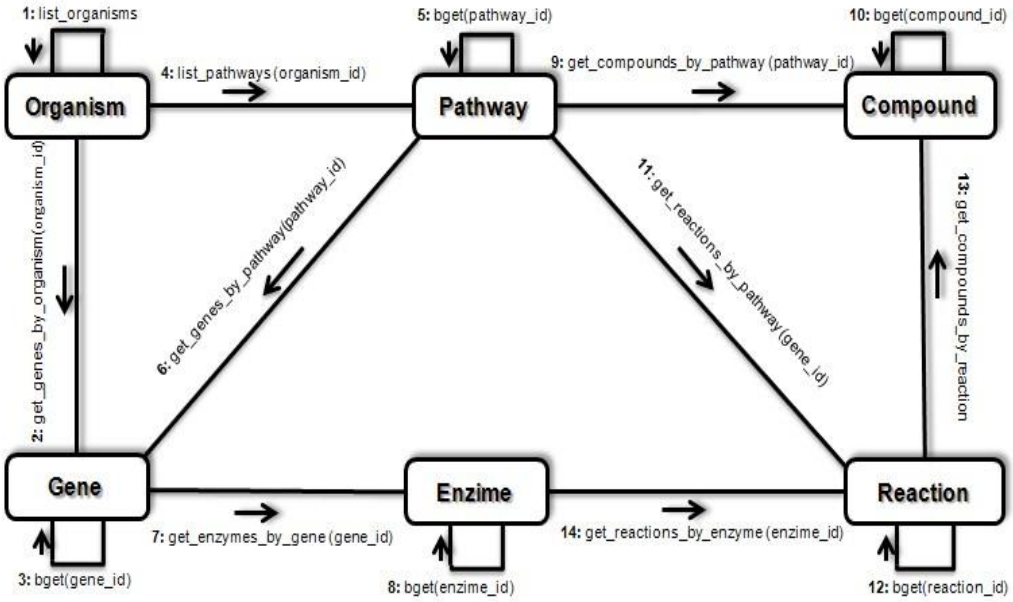

Figura 5. Representa los métodos utilizados del WS de KEGG y el orden de descarga de la información.

2.3 Herramientas utilizadas para la construcción del SWC

2.3.1 Tecnología Java

Entre sus objetivos principales se encuentra proveer un lenguaje relativamente fácil de usar, ya que fue diseñado con la idea de eliminar muchas de las fallas de otros lenguajes. Está orientado a objetos, habilita a los usuarios para crear código claro y racional, provee un entorno interpretado para aumentar la velocidad de desarrollo, además de proporcionar portabilidad en el código generado [Deitel and Deitel, (1999)]. Otras de las características de la tecnología Java es que permite a los usuarios ejecutar más de un hilo de actividad, carga clases a memoria de manera dinámica cuando estas se necesitan, soporta cambios de programa de manera dinámica durante la ejecución cargando clases desde diferentes fuentes y proporciona una mejor seguridad en la ejecución de código.

2.3.2 Lenguaje Java

El lenguaje de programación Java es un lenguaje de programación de alto nivel y orientado a objeto el cual se basa en los lenguajes C y C++, pero es más fácil de utilizar y elimina herramientas de bajo nivel como la manipulación de punteros. Se utiliza para desarrollar tanto *applets* (aplicaciones que se integran a una página web y son ejecutadas por un navegador) como aplicaciones de escritorio. Entre las

principales características de este lenguaje se encuentran las siguientes [Horstmann, (2010)]:

1) Lenguaje simple: elimina la complejidad de los lenguajes como C y da paso al contexto de los lenguajes modernos orientados a objetos.
2) Orientado a Objetos: fue creado desde sus inicios como un lenguaje orientado a objeto, utilizando clases, las cuales poseen datos y funciones encargadas de manejar estos datos.
3) Distribuido: proporciona una colección de clases para su uso en aplicaciones de red, que permiten establecer y aceptar conexiones con servidores o clientes remotos, facilitando así la creación de aplicaciones distribuidas.
4) Interpretado y compilado a la vez: Java es compilado, en la medida en que su código fuente se transforma en una especie de código máquina, los bytecodes[1], semejantes a las instrucciones de ensamblador. Por otra parte, es interpretado, pues los bytecodes se pueden ejecutar directamente sobre cualquier máquina a la cual se hayan portado el intérprete y el sistema de ejecución en tiempo real.
5) Robusto: Java proporciona un gran número de comprobaciones en compilación y en tiempo de ejecución, por lo que fue creado para obtener aplicaciones altamente confiables.
6) Indiferente a la arquitectura: Java está diseñado para soportar aplicaciones que serán ejecutadas en los más variados entornos de red, desde Unix a Windows Nt, pasando por Mac y estaciones de trabajo, sobre arquitecturas distintas y con sistemas operativos diversos. Al compilar un programa en Java, el bytecode resultante es interpretado por diferentes computadoras de igual manera, solamente hay que implementar un intérprete para cada plataforma. De esa manera Java logra ser un lenguaje que no depende de una arquitectura computacional definida.
7) Portable: la indiferencia a la arquitectura representa solo una parte de su portabilidad. Además, Java especifica los tamaños de sus tipos de datos básicos y el comportamiento de sus operadores aritméticos, de manera que los programas son iguales en todas las plataformas.

[1]Bytecode: código compilado de Java

2.3.3 Plataforma Java

Con plataforma se hace referencia al ambiente de hardware y software en donde el programa se ejecuta, por ejemplo, plataformas como Linux, Solaris, Windows 2003 y MacOS. En casi todos los casos las plataformas son descritas como la combinación del sistema operativo y el hardware. La plataforma Java se diferencia de estas plataformas en que es una plataforma solo de software y se ejecuta sobre las otras plataformas de hardware. Está compuesta por la Máquina Virtual de Java (JVM, del inglés *Java Virtual Machine*) y la Interfaz de Programación de Aplicaciones (API, del inglés *Application Programming Interface*) de Java [Gosling et al. (2005)]. La JVM es una de las piezas fundamentales de la plataforma Java, que tiene como principal ventaja la de aportar portabilidad al lenguaje. Básicamente se sitúa en un nivel superior al hardware del sistema sobre el que se pretende ejecutar la aplicación y este actúa como un puente que entiende tanto el bytecode, como el sistema sobre el que se pretende ejecutar. Así, cuando se escribe una aplicación Java, se hace pensando que será ejecutada en una máquina virtual Java en concreto, siendo esta la que en última instancia convierte de código bytecode a código nativo del dispositivo final.

La API Java está provista por los creadores del lenguaje Java, y que da a los programadores los medios para desarrollar aplicaciones Java. Como el lenguaje Java es un lenguaje orientado a objetos, la API de Java provee un conjunto de clases utilitarias para efectuar toda clase de tareas necesarias dentro de un programa, brindando una gran colección de componentes de software que proporcionan muchas utilidades para el programador, por ejemplo, las API's para las interfaces gráficas. La API Java está organizada en paquetes lógicos, donde cada paquete contiene un conjunto de clases relacionadas semánticamente.

2.3.4 Entorno de desarrollo integrado. NetBeans IDE 6.8

Un entorno de desarrollo integrado (IDE, del inglés *Integrated Development Environment*) es un programa compuesto por un conjunto de herramientas para un programador. Puede dedicarse en exclusiva a un solo lenguaje de programación o puede utilizarse para varios. Un IDE es un entorno de programación que ha sido empaquetado como un programa de aplicación, es decir, consiste en un editor de código, un compilador, un depurador y un constructor de interfaz gráfica [Myatt, (2008)].

El NetBeans IDE es un entorno de desarrollo, una herramienta para programadores pensada para escribir, compilar, depurar y ejecutar programas. Está escrito en Java pero puede servir para cualquier otro lenguaje de programación. Existe además un número importante de módulos para extender el IDE NetBeans. El IDE NetBeans es un producto libre y gratuito sin restricciones de uso.

NetBeans IDE 6.8 se ejecuta en muchas plataformas incluyendo Windows, Linux, Mac OS X y Solaris; y posee un proceso simplificado de instalación que permite instalar y configurar fácilmente el IDE para satisfacer exactamente varias necesidades. Tiene muy buenas opciones nuevas, y cada vez más sus desarrolladores agregan nuevas características y capacidades para otros lenguajes.

2.3.5 Sistema Gestor de base de datos

Un SGBD facilita las tareas de administración de los datos y acelera el desarrollo de la aplicación, por lo que se hace necesario realizar una selección adecuada. En la actualidad existe una gran variedad de SGBD, tanto de tipo comercial como libre, entre los cuales se encuentra Microsoft SQL Server, Oracle, Microsoft Access, MySQL, PostgreSQL entre otros. Para la selección del gestor de base de datos se tuvieron en cuenta MySQL y PostgreSQL ya que son sistemas de gestión de base de datos relacional y de fuente abierta.

Principales características de PostgreSQL

PostgreSQL es un sistema de gestión de bases de datos objeto-relacional basado en el proyecto POSTGRES, de la universidad de Berkeley. Es un sistema objeto-relacional, ya que incluye características de la orientación a objetos, como puede ser la herencia, tipos de datos, funciones, restricciones, disparadores, reglas e integridad transaccional. Está diseñado para ambientes de grandes volúmenes de información, usando una estrategia de almacenamiento de filas llamada Acceso Concurrente Multiversión (MVCC, del inglés *Multiversion Concurrency Control*) para conseguir una mejor respuesta en ambientes de grandes volúmenes; lo cual permite que mientras un proceso escribe en una tabla, otros accedan a la misma sin necesidad de bloqueo. Cumple completamente con ACID (del inglés, *Atomicity, Consistency, Isolation, Durability*) y con el Instituto Nacional Americano de Estándares ANSISQL (del inglés, *American National Standards Institute*). Su avanzada funcionalidad se pone de manifiesto con las consultas SQL declarativas, soporte multiusuario, transacciones, optimización de consultas, herencia y valores no atómicos (atributos basados en

vectores y conjuntos). Cuenta con una API flexible lo cual ha permitido dar soporte para el desarrollo con PostgreSQL en diversos lenguajes de programación.

Una de las características de PostgreSQL en que aventaja a la base de datos MySQL es que permite al usuario o administrador de la base de datos escribir sus propias funciones dentro de la base de datos; estas funciones se almacenan y ejecutan desde el proceso de base de datos y no desde la aplicación del cliente [Smith, (2010)].

Después de haber realizado un estudio exhaustivo de las principales características de este gestor de bases de datos, se puede llegar a la conclusión de que PostgreSQL ofrece una garantía de integridad de los datos mucho más fuerte que otros gestores, por lo que en aquellos escenarios en los que no puede permitirse que se corrompa o se pierda ni un solo registro, solo es una opción utilizar PostgreSQL. Aunque sea más lento respondiendo a una única consulta, este presenta una mejor estabilidad y rendimiento bajo grandes cargas de trabajo.

Teniendo en cuenta estas potencialidades y dado que se necesitará de un manejo complejo de grandes volúmenes de información se seleccionó PostgreSQL como SGBD.

2.4 Ingeniería de software del SWC

2.4.1 Requisitos Funcionales del SWC

Los requisitos funcionales son las capacidades o condiciones que el sistema debe cumplir; en este caso se plantearon los siguientes:

R1. Gestionar organismo.
R2. Gestionar genes por organismo.
R3. Gestionar *pathway* por organismo.
R4. Gestionar genes por *pathway*.
R5. Gestionar compuestos por *pathway*.
R6. Gestionar reacciones por *pathway*.
R7. Gestionar compuestos por reacciones.
R8. Gestionar enzimas por gen.
R9. Gestionar reacciones por enzima.
R10. Gestionar compuesto por enzima.
R11. Configurar la conexión con la base de datos.

R12. Configurar conexión a Internet.

R13. Importar base de datos.

R14. Exportar base de datos.

R15. Seleccionar organismos a descargar.

R16. Mostrar ayuda de la aplicación.

2.4.2 Requisitos no Funcionales del SWC

Los requerimientos no funcionales son propiedades o cualidades que el producto debe tener. Para el desarrollo de la aplicación es indispensable tener en cuenta los siguientes requisitos no funcionales:

1. Requisitos de apariencia: el sistema deberá tener una interfaz externa amigable, que sea sencilla y fácil de manipular por el usuario.
2. Requisitos de usabilidad: la aplicación resulta de fácil uso para personas sin experiencia previa con las computadoras.
3. Requisitos de software: para el funcionamiento de este software deberá estar instalada la máquina virtual de Java y el SGBD usado en PostgreSQL.
4. Requisitos de hardware: se debe contar con 256 MB de memoria RAM como mínimo, aunque lo ideal sería 2 GB. Es necesaria la implementación de los dispositivos de conexión necesarios como MODEM o Red LAN con acceso a Internet.
5. Requisitos de diseño e implementación: es necesario tener instalado PostgreSQL, NetBeans IDE 6.8 como herramienta de desarrollo y Enterprise Architect 7.5.
6. Documentación: el usuario podrá auxiliarse de una ayuda del sistema en todo momento, para lograr un fácil uso del mismo.

2.4.3 Análisis de la aplicación

Rol del sistema

Tabla 1. Descripción del actor para el WSC.

Actor	Justificación
Usuario	Es la persona que va a interactuar con el sistema para realizar la descarga de la información biológica.

2.4.4 Modelo de Casos de Uso

Un caso de uso determina un grupo de acciones secuenciales que el sistema puede llevar a cabo a través de sus actores, incluyendo alternativas dentro de la secuencia; es decir que constituye un fragmento de funcionalidad que el sistema ofrece para aportar un resultado de valor para sus actores. Partiendo de los requerimientos definidos y la relación con el actor del sistema se obtuvo el siguiente diagrama de casos de uso.

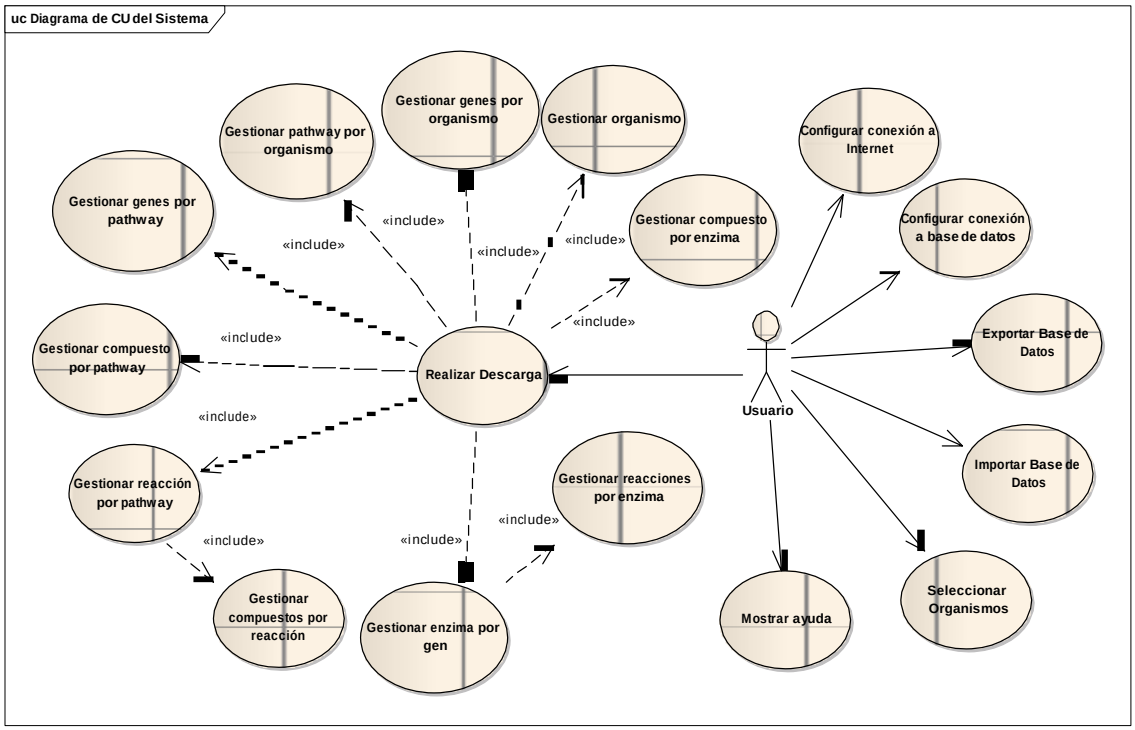

Figura 6. Diagrama de Casos de Uso de WSC.

2.4.5 Diagramas de clases del Análisis y el Diseño

Un diagrama de clases es un tipo de diagrama estático que describe la estructura de un sistema mostrando sus clases, atributos y las relaciones entre ellos. Los diagramas de clases son utilizados durante el proceso de análisis y diseño de los sistemas, donde se crea el diseño conceptual de la información que se manejará en el sistema, los componentes que se encargarán del funcionamiento y la relación entre los diferentes elementos de ambos procesos.

A continuación se representan los diagramas de clases del Análisis y el Diseño para los Casos de Uso Configurar Conexión a Internet y Realizar Descarga.

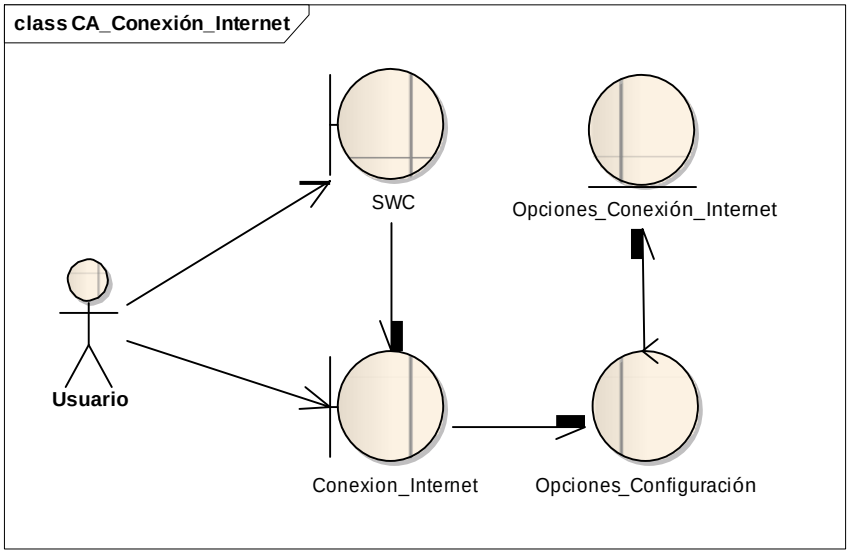

Figura 7. Diagrama de Clases del Análisis.CU_ Configurar Conexión a Internet.

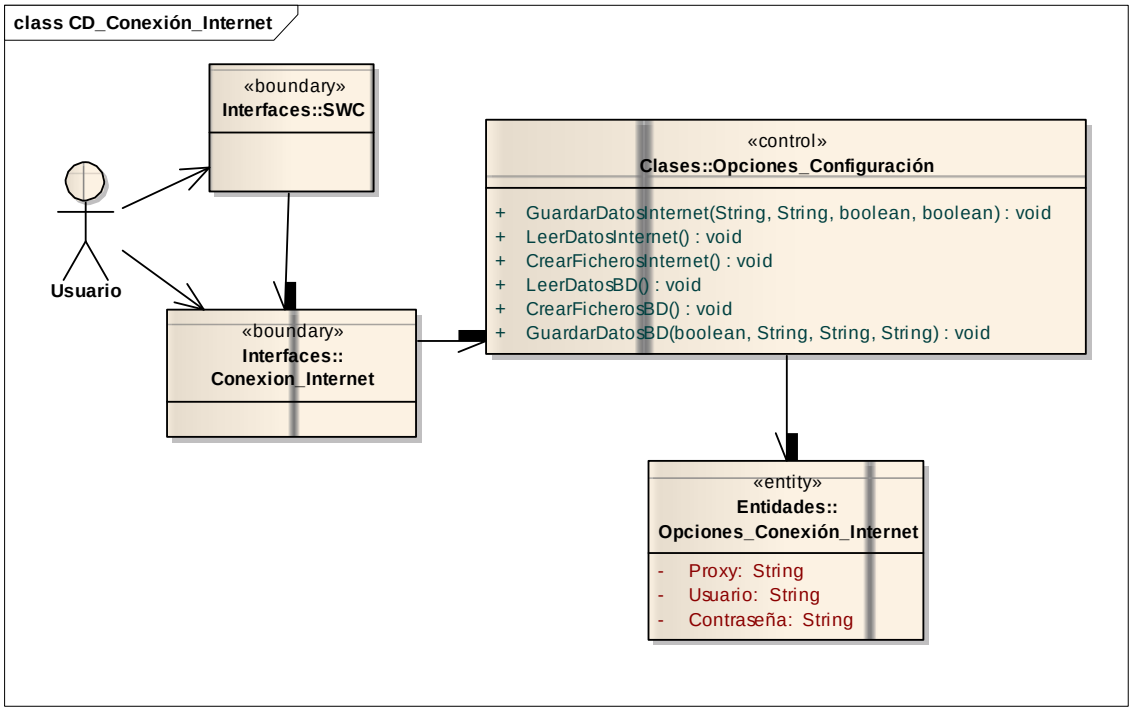

Figura 8. Diagrama de Clases del Diseño.CU_ Configurar Conexión a Internet.

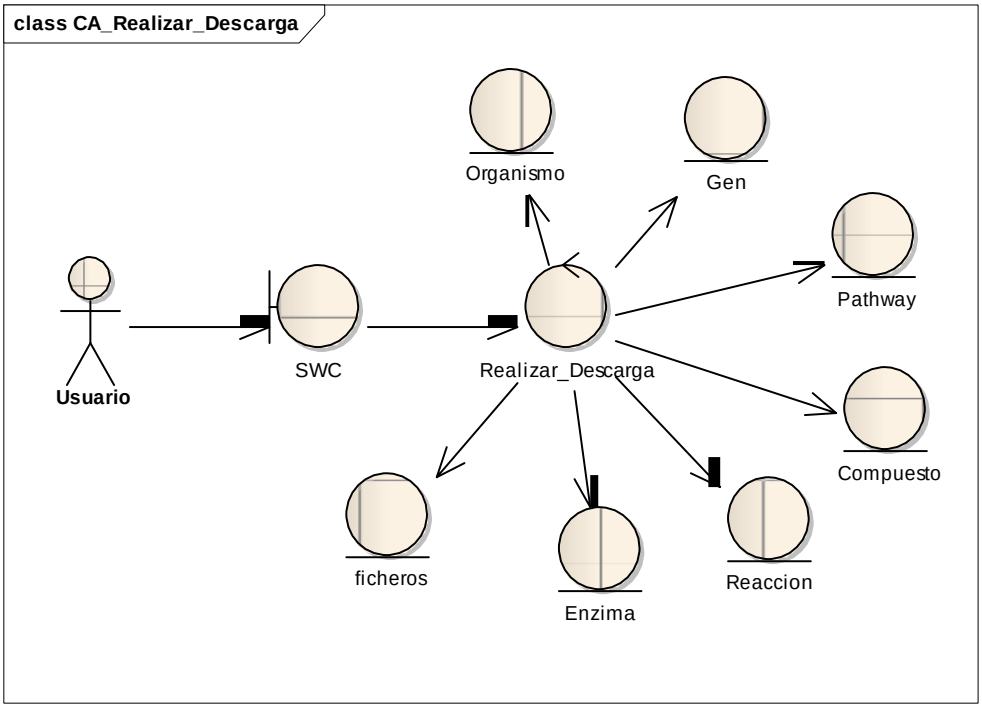

Figura 9. Diagrama de Clases del Análisis. CU_Realizar Descarga.

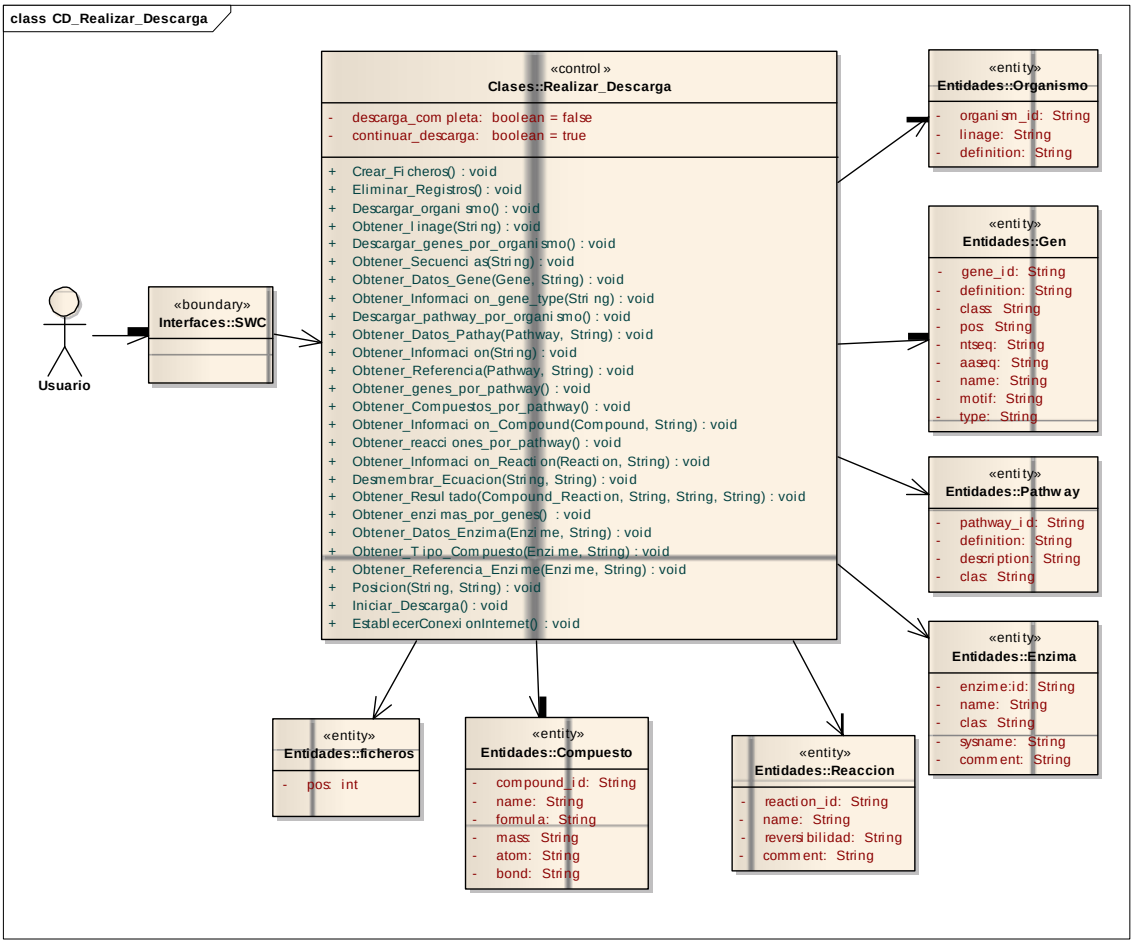

Figura 10. Diagrama de Clases del Diseño. CU_Realizar Descarga.

2.4.6 Modelo de implementación

El modelo de implementación describe cómo se organizan los componentes de acuerdo con el lenguaje de programación utilizado y al entorno de implementación y cómo dependen los componentes entre sí. Un componente es el empaquetamiento físico de un elemento del diseño, como lo son las clases en el modelo de diseño. Como se puede observar, este diagrama se utiliza para modelar la vista estática de un sistema. Muestra la organización y las dependencias lógicas entre un conjunto de componentes software, sean estos componentes de código fuente, tablas de la base de datos o ejecutables. En este caso específico se relaciona el Ejecutable con el resto de los componentes que maneja la aplicación para el desarrollo de la misma.

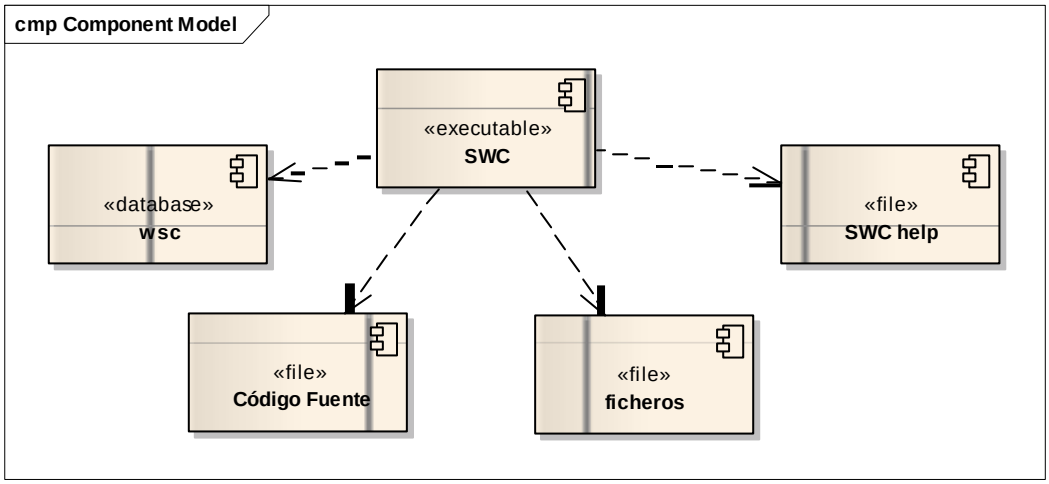

Figura 11. Modelo de componentes del WSC.

2.4.7 Modelo de Datos

El modelo de datos es el artefacto resultante de la actividad de diseño de la base de datos; este describe la representación lógica y física de los datos persistentes. A continuación se muestran las entidades de la base de datos, con sus atributos y las relaciones que existen entre ellas.

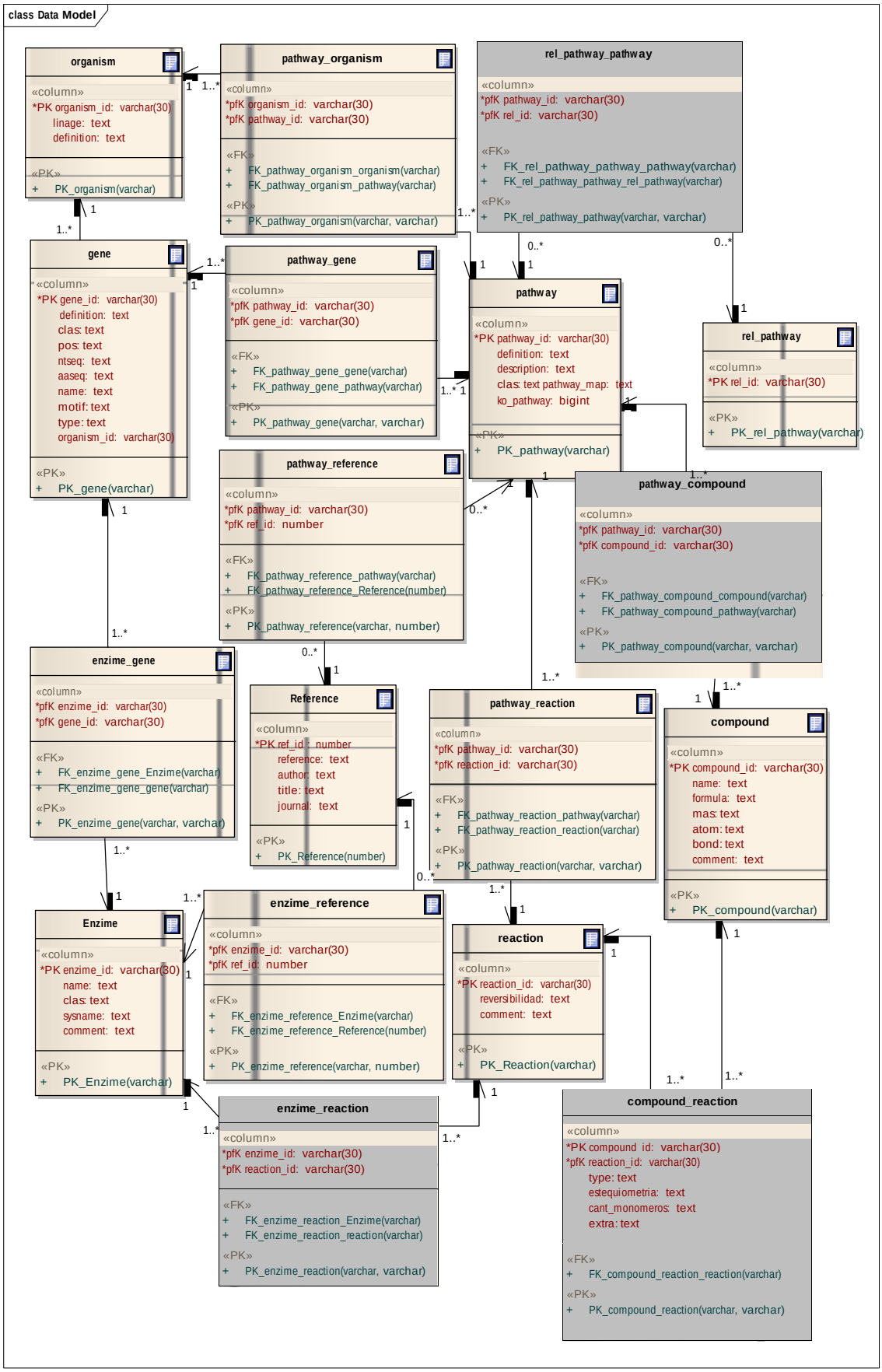

Figura 12. Modelo de datos.

2.5 Deficiencias en la información biológica

Una vez diseñada y construida nuestra base de datos a partir de KEGG, se pudo constatar en la información obtenida algunas deficiencias, ya sea relacionada con la **no completitud** o **nomenclatura no unívoca** de la información [Reyes et al. (En preparación)].En el caso de la no completitud de la información se encuentran los genes y las reacciones metabólicas. En los genes, una vez utilizado el método *bget (gene_id)* para obtener los datos de los mismos, aparecen casos en los que el elemento *name* no se obtiene a partir del WS de KEGG. Esto constituía un grave problema a la hora de reconstruir los modelos metabólicos a escala genómica, si se tiene en cuenta que dichos modelos se reconstruyen a partir de la anotación del genoma. Otro ejemplo de no completitud es el caso de la reversibilidad dentro de las reacciones metabólicas, donde una vez utilizado el método *bget (reaction_id)* para obtener todos los datos de las reacciones, se pudo comprobar que el WS de KEGG devuelve siempre el elemento *type* como Reversible, a pesar de estar demostrado que existen reacciones donde este parámetro se comporta como Irreversible. La reversibilidad de las reacciones es otro elemento clave en la reconstrucción de modelos metabólicos a escala genómica, si se tiene en cuenta que cada modelo se puede representar como una red, donde los metabolitos sustratos se conectan con los metabolitos productos mediante conexiones dirigidas o no, en dependencia del tipo de reversibilidad. Finalmente, estas redes son estudiadas para un mejor análisis de los sistemas biológicos.

Otra de las deficiencias estuvo relacionada con la nomenclatura no unívoca de los compuestos. La mayoría de las bases de datos biológicas no siguen un estándar para definir el nombre de los mismos. Por el contrario, cada grupo de investigación a la hora de reconstruir un modelo metabólico de un organismo asume un nombre diferente para los compuestos. En el mejor de los casos ponen a disposición de la comunidad científica una homología entre los nombres descritos por ellos y sus correspondientes en KEGG. Esto representaba otra dificultad debido a que los compuestos son una pieza clave dentro de la reconstrucción de un modelo metabólico. Igualmente, en el caso de KEGG se asumen diferentes nombres para un mismo compuesto, por lo que era necesario contrastar con otras bases de datos específicas de compuestos para asociar a cada uno el nombre que más lo identifica dentro de la comunidad científica.

2.6 Métodos de integración de bases de datos biológicas

Estas deficiencias encontradas en nuestra base de datos nos hicieron buscar nuevas alternativas a través de la integración con otras bases de datos específicas para genes y compuestos [Reyes et al. (En preparación)]. A continuación se explica la solución.

2.6.1 Completamiento de la información en genes y reacciones

En el primer caso era necesario completar la información de los genes, a partir de una base de datos específica para los mismos. La solución partió de integrar nuestra base de datos con NCBIGeneID (NCBI, del inglés *National Center for Biotechnology Information*). Esta base de datos hospeda secuencias genómicas y otras informaciones de interés biológico en GenBank [Miller et al. (2009)], habilitadas a través de Entrez [NCBI, (2012)], un motor de búsqueda online. NCBIGeneID proporciona un WS que entre sus funcionalidades está *getGeneName()*, lo que facilita completar el nombre de los genes que tenemos en nuestra base de datos (Ver Figura 13). Este método interpreta correctamente el identificador de KEGG para los genes, con lo cual eso no representa incongruencias a la hora de obtener la información desde NCBIGeneID.

En el segundo caso pudimos comprobar que el ftp de KEGG proporcionaba la reversibilidad de las reacciones metabólicas correctamente. Se definió el método *getReactionType(),* donde a partir del listado de pathways para un organismo se obtiene el listado de reacciones metabólicas que conforman cada *pathway* en formato KGML(del inglés *KEGG Markup Language*). Este fichero tiene un parámetro *Type* que define el tipo de reacción: Reversible o Irreversible, que es utilizado para corregir la información en la base de datos de COPABI (Ver Figura 13).

2.6.2 Nomenclatura unívoca en la definición de los compuestos

Como ya habíamos mencionado los compuestos representan una de las principales deficiencias encontradas en la base de datos, si se tiene en cuenta que cada uno presenta varios nombres y los grupos de investigación no llegan a un consenso en cuanto a tomar un único nombre para un mismo compuesto. En este sentido decidimos integrar la base de datos con CHEBI, una colección específica de compuestos químicos que proporciona un vocabulario estandarizado y sin ambigüedades en su terminología [Degtyarenko et al. (2008)]. Igualmente, esta base de datos provee por cada compuesto varios nombres, aunque el primero de ellos se

corresponde con el que más identifica a ese compuesto, según *International Union of Pure and Applied Chemistry* (IUPAC) and *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology* (NC-IUBMB).

Sin embargo, se planteaba el problema de que CHEBI no interpreta el identificador de KEGG, por ejemplo: C00010, y para lo cual fue necesario utilizar SABIO-RK [Wittig, (2009)], una base de datos transitoria que permitió convertir el identificador de KEGG al de CHEBI a través del WS que provee. El método utilizado desde SABIO-RK fue *getCompoundIDFromKEGGID()*. Una vez obtenido el identificador de CHEBI se obtiene a través del WS de esta base de datos el nombre del compuesto. El método que se utiliza en este caso es el *getCompoundFromChebi()*. El siguiente diagrama representa la forma en que se obtiene la información a partir de la integración con diversas bases de datos biológicas.

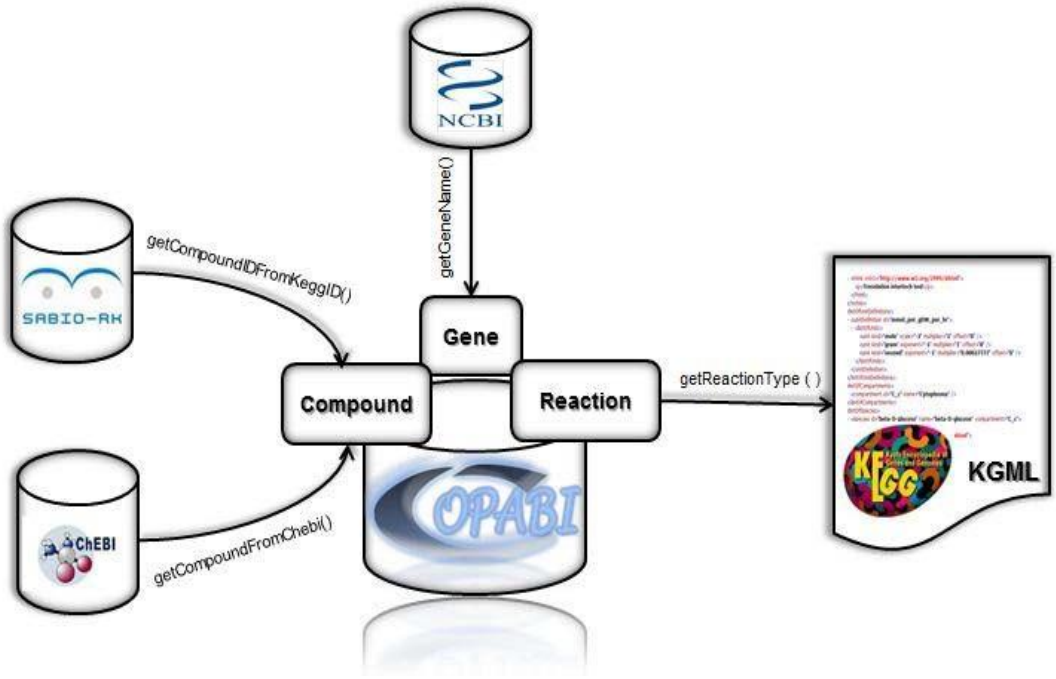

Figura 13. Integración con bases de datos biológicas.

Capítulo 3

Implementación de COPABI

Todas las bases de datos biológicas disponibles en internet cuentan con una aplicación web para mostrar la información que ellas poseen, así como métodos para exportar dicha información en diferentes formatos. Como se pudo apreciar en el capítulo anterior, nos basamos en KEGG para obtener la información biológica necesaria para la reconstrucción de modelos metabólicos a escala genómica, almacenándola en una base de datos. Sin embargo, se hacía necesario manipular dicha información en un entorno integrado que permitiese además reconstruir los modelos metabólicos a escala genómica de los organismos, y es así como surge COPABI.

En este capítulo se analizarán las herramientas y tecnologías aplicadas para el diseño e implementación de COPABI, así como diferentes formatos para exportar dicha información. Por último se reflejará la Ingeniería de Software enfocada a dicha aplicación web.

3.1 Herramientas CASE. Enterprise Architect

Las herramientas de ingeniería de software asistida por computador (CASE, del inglés *Computer Aided Software Engineering)* son diversas aplicaciones informáticas que permiten aumentar la productividad en el desarrollo de software, reduciendo el coste en términos de tiempo y de dinero. En la actualidad existe un gran número de estas aplicaciones como son Rational Rose, Visual Paradigm, Enterprise Architect, entre otras, las cuales se utilizan para la creación de artefactos durante el desarrollo de un software [Sparx Systems, (2011)]. Se ha decidido utilizar Enterprise Architect, pues proporciona un entorno de modelación de carácter colaborativo y potenciado mediante el Lenguaje Unificado de Modelado (UML, del inglés *Unified Modeling Language*) [Booch et al. (2000)]. Es una herramienta multi-usuario, basada en Windows, diseñada para ayudar a construir software robusto y fácil de mantener.

3.2 Servidor Web

Un servidor Web es un programa que se ejecuta continuamente en un ordenador, manteniéndose a la espera de peticiones por parte de un cliente. Responde a las mismas adecuadamente mediante una página Web que se exhibirá en el navegador o mostrando el respectivo mensaje si se detectó algún error [Yeager and McGrath, (1996)].

3.2.1 Apache

Apache es el servidor Web más utilizado del mundo. Es un software de código abierto que funciona sobre cualquier plataforma. Desde su origen ha evolucionado hasta convertirse en uno de los mejores servidores en términos de eficiencia, funcionalidad y velocidad [Tong, (2008)].

Ventajas de utilizar Apache

- Es flexible y extensible, dando la posibilidad de ampliar sus capacidades y funcionalidades mediante módulos.
- Es altamente fiable pues aproximadamente el 90% de los servidores con más alta disponibilidad funcionan con él.
- Se destaca por su gran velocidad.

3.3 Aplicación Web

Una aplicación web es una solución informática que los usuarios utilizan accediendo a un servidor web a través de Internet o de una intranet. Aplicaciones como los webmails, wikis y weblogs son ejemplos bien conocidos de aplicaciones web.

Ventajas de utilizar una aplicación web

Es necesaria una aplicación Web que funcione como interfaz del sistema que se propone. Son varios los argumentos a favor de esta opción, entre ellos se destacan:

- Solo requiere del uso de un navegador web.
- Son independientes del sistema operativo del usuario final.
- Habilidad para actualizar y mantener aplicaciones web sin distribuir e instalar software en miles de clientes, lo que significa una reducción sensible de costo y tiempo.

3.4 Lenguajes utilizados en la implementación de COPABI

Enfrascados en la tarea de seleccionar las herramientas a usar para la implementación de COPABI, se tuvo en cuenta que en el caso de las aplicaciones web las opciones a escoger se dividen en dos grupos: los lenguajes que se ejecutan del lado del servidor, entre los que está PHP (del inglés *Hypertext Preprocessor*) y los que corren del lado del cliente como HTML, Javascript, etc. A continuación se caracterizan los lenguajes usados para la implementación de COPABI.

3.4.1 HTML Y XHTML

HTML fue creado en 1986 por el físico nuclear Tim Berners-Lee. Es un lenguaje muy sencillo que permite describir hipertexto, es decir, texto presentado de forma estructurada y agradable, utilizado normalmente en la WWW. Contiene enlaces que conducen a otros documentos o fuentes de información relacionadas, y con inserciones multimedia (gráficos, sonido, etc). Una de las características esenciales de este lenguaje es la universalidad, lo que significa que prácticamente cualquier ordenador, independientemente del sistema operativo, puede leer o interpretar una página web [Castro, (2006)].

El Lenguaje extensible de marcado de hipertexto (XHTML, del inglés *Extensible Hypertext Markup Language*), es la versión XML de HTML. Este nuevo lenguaje tiene todas las características de HTML y por tanto lo pueden entender todos los navegadores. Como utiliza la sintaxis de XML, gana toda la potencia y flexibilidad de XML y es una base perfecta para las Hojas de Estilo en Cascada (CSS, del inglés *Cascading Style Sheets*) [Castro, (2006)].

3.4.2 JavaScript

JavaScript es un lenguaje orientado a objetos. Los programas JavaScript son ficheros textos con código estándar para el intercambio de información (ASCII, del inglés, *American Standard Code for Information Interchange)*. Pueden ser incluidos en ficheros aparte o en la misma página HTML y viajar así al cliente, permitiendo prestar interactividad a las páginas, así como para las validaciones de los datos que se introducen en la aplicación [Stefanov, (2010)]. Los archivos de tipo JavaScript son documentos normales de texto con la extensión .js, que se pueden crear con cualquier editor de texto como Notepad, Wordpad, etc.

JavaScript es un lenguaje que simplifica el código XHTML de la página, además de que se puede reutilizar el mismo código JavaScript en todas las páginas del sitio web y

que cualquier modificación realizada en el archivo se ve reflejada inmediatamente en todas las páginas XHTML que lo enlazan.

3.4.3 PHP

PHP es un lenguaje de código abierto especialmente adecuado para desarrollo web y que puede ser incrustado en HTML. Lo que distingue a PHP es que el código es ejecutado en el servidor, generando HTML y enviándolo al cliente. El cliente recibirá los resultados de ejecutar el script, sin ninguna posibilidad de determinar qué código ha producido el resultado recibido. El servidor web puede ser incluso configurado para que procese todos los archivos HTML con PHP. Lo mejor de usar PHP es que es extremadamente simple para el principiante, pero a su vez, ofrece muchas características avanzadas para los programadores profesionales [Schlossnagle, (2007)].

Otra de las principales ventajas que ofrece PHP es ser un lenguaje libre y abierto, pues su código fuente está disponible y es gratuito. Inicialmente esta tecnología fue diseñada para entornos UNIX por lo que ofrece más prestaciones en este sistema operativo, pero es perfectamente compatible con Windows [Gutmans et al. (2004)].

3.4.4 Framework CodeIgniter 1.7.3

Un framework, (WAF, del inglés *Web Application Framework*), es una serie de librerías y clases que se han unido bajo un único esquema de colaboración para lograr el desarrollo rápido de aplicaciones (RAD, del inglés *Rapid Application Development*) [Schlossnagle, (2007)].

Su esencia consiste en que simplifica y acelera considerablemente el proceso de desarrollo de una aplicación, ya que automatiza algunos de los patrones utilizados para resolver las tareas más comunes, mediante el encapsulamiento de operaciones complejas en instrucciones sencillas. Todas estas ventajas hicieron irrevocable la decisión de utilizar un framework para el desarrollo de la solución de software, pues la reutilización de código y otras características permiten al desarrollador dedicarse por completo a los aspectos específicos de la aplicación en cuestión.

CodeIgniter es un framework para construir sitios web usando PHP. Su objetivo es habilitar el desarrollo de proyectos mucho más rápido de lo que se podría si se escribiese código desde cero, a través de un conjunto de librerías para tareas comúnmente necesarias, tanto como una simple interfaz y estructura lógica para

acceder a estas librerías. CodeIgniter permite concentrarse creativamente en el proyecto minimizando el volumen de código necesario para una tarea determinada [Upton, (2007)].

3.4.5 Modelo-Vista-Controlador

CodeIgniter está basado en el patrón de desarrollo Modelo-Vista-Controlador (MVC). MVC es una aproximación al software que separa la lógica de la aplicación de la presentación. En la práctica, permite que sus páginas web contengan mínima codificación ya que la presentación es separada del código PHP.

1) El Modelo representa la estructura de datos. Típicamente sus clases contendrán funciones que lo ayudarán a recuperar, insertar y actualizar información de la base de datos.
2) La Vista es la información que es presentada al usuario. Normalmente será una página web, pero en CodeIgniter, una vista también puede ser un fragmento de una página como un encabezado o un pie de página.
3) El Controlador sirve como un intermediario entre el Modelo, la Vista y cualquier otro recurso necesario para procesar la petición HTTP y generar una página web.

3.5 Formatos para exportar la información biológica

Este es un aspecto primordial a la hora de brindar la información biológica almacenada en la base de datos y que puede influir de forma determinante en el uso de la misma para un posterior análisis con otras herramientas implementadas en el campo de la Biología de Sistemas. En este sentido se puede decir que las bases de datos mencionadas en el capítulo anterior utilizan diferentes formatos a la hora de mostrar la información, los más utilizados son:

- Lenguaje de Marcado para la Biología de Sistemas (SBML, del inglés *Systems Biology Markup Language*). Es el nombre de un lenguaje de descripción basado en XML que se utiliza para representar modelos de procesos biológicos. SBML puede representar redes metabólicas, rutas de señalización celular, redes de regulación génicas y muchas otras clases de sistemas [Hucka et al. (2003)].
- FASTA. En Bioinformática es un formato de fichero informático basado en texto, utilizado para representar secuencias ya sean de ácidos nucleicos o de

péptidos, y en el que los pares de bases o los aminoácidos se representan usando códigos de una única letra [Books, (2010)].

- BLAST, (del inglés, *Basic Local Alignment Search Tool*). Es un programa informático de alineamiento de secuencias de tipo local, ya sea de ADN o de proteínas [Camacho et al. (2009)].
- KGML. Es un lenguaje basado en XML pero que define una estructura propia para la base de datos japonesa [Kanehisa and Goto, (2010)].

En el caso de COPABI se decidió utilizar dos formas de exportar la información biológica: 1) el estándar SBML nivel 2 versión 1 y 2) en el caso específico de los modelos metabólicos, además de exportarlos en el formato anterior, se exportan también siguiendo los requerimientos de entrada del OptGene, un software que permite obtener la mejor combinación de supresión de genes para la optimización de una función objetivo fenotípica deseada en un sistema biológico determinado [Patil et al.(2005)], también llamado BioOpt y usado en BioMet toolbox (www.sysbio.se/biomet).

3.6 Ingeniería de software de COPABI

3.6.1 Modelo de dominio

El modelo de dominio es una representación visual de los conceptos u objetos del mundo real significativos para un problema o área de interés. Este es utilizado para comprender, capturar y describir los conceptos más importantes empleados en el contexto del negocio. Para la construcción de un modelo de dominio se extraen los conceptos y eventos principales del entorno y se relacionan en un diagrama de clases usando UML (Ver Figura 14).

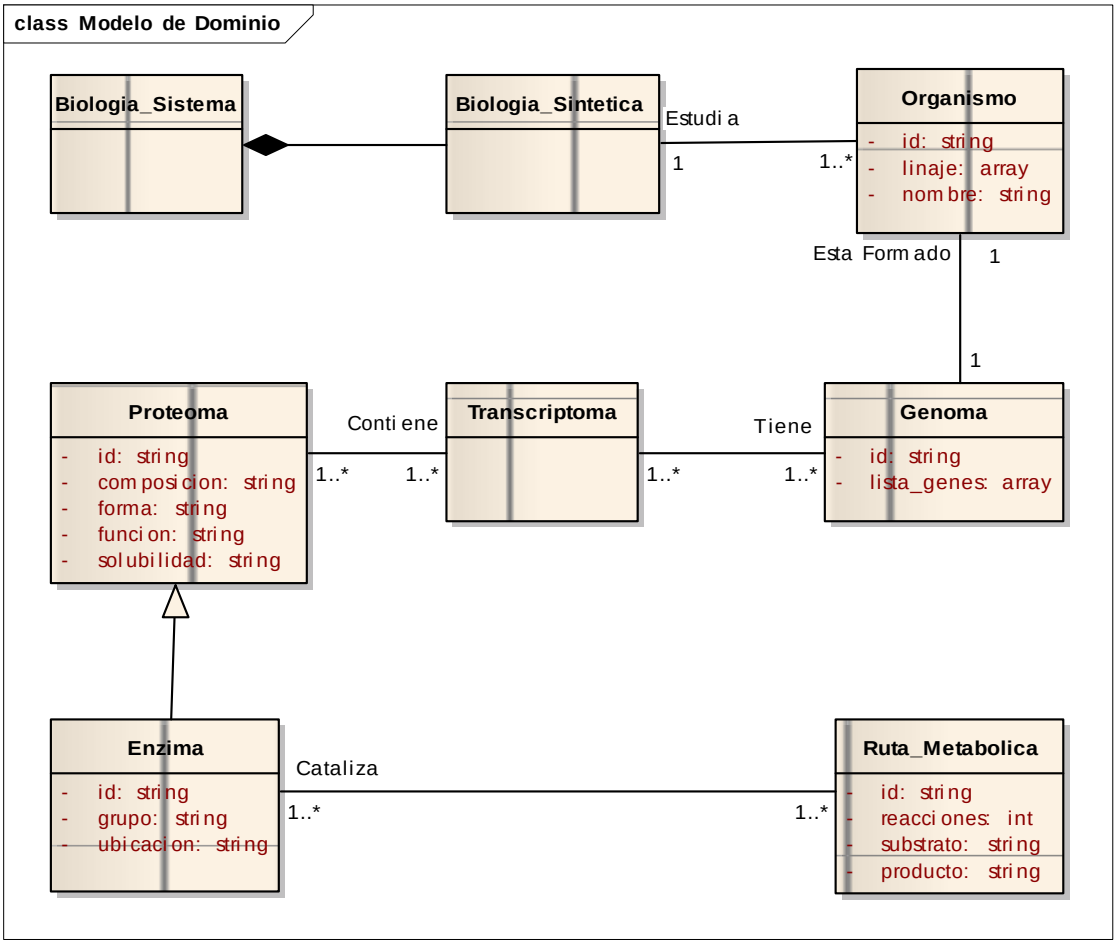

Figura 14. Modelo de Dominio.

Conceptos que se utilizan en el modelo de dominio

Para la realización del modelo de dominio es necesario identificar los conceptos principales del negocio ya que esto permite un mejor entendimiento del objeto de estudio y facilita la captura de los requisitos; posibilitando el desarrollo de un software de acuerdo a las necesidades de los clientes. A continuación se presentan los conceptos fundamentales del modelo de dominio:

- Biología de Sistema: área de investigación que se remonta a la década del 60 del siglo pasado, pero que tuvo un auge a partir del año 2000. Se encarga de estudiar todas las interacciones que se producen dentro de los sistemas biológicos, vistos desde un enfoque sistémico. Para esto usa herramientas de simulación, modelación y comparación.
- Biología Sintética: se basa en el diseño y construcción de sistemas biológicos artificiales con nuevas funcionalidades, basándose en una metodología

ingenieril, lo que implica la interrelación de varias ciencias como Informática, Matemática, Física, entre otras.

- Organismo: Entidad biológica capaz de reproducirse o de transferir material genético, incluyéndose dentro de este concepto a las entidades microbiológicas, sean o no celulares. Casi todo organismo está formado por células, que pueden agruparse en órganos, y estos a su vez en sistemas, cada uno de los cuales realizan funciones específicas.
- Genoma: conjunto de la información genética, codificada en una o varias moléculas de ADN (en muy pocas especies en Ácido ribonucleico (ARN)), donde están almacenadas las claves para la diferenciación de las células que forman los diferentes tejidos y órganos de un individuo.
- Transcriptoma: es el conjunto de genes que se están expresando en un momento dado en una célula. La expresión de un gen supone que este ha sido transcrito a ARN mensajero. Células de un mismo organismo y con un mismo genoma pueden llegar a ser tipos celulares muy dispares dependiendo de la combinación de genes que exprese cada una, o lo que es lo mismo, dependiendo de su transcriptoma.
- Proteoma: es la totalidad de proteínas expresadas en una célula particular bajo condiciones de medioambiente y etapa de desarrollo, (o ciclo celular) específicas, como lo puede ser la exposición a estimulación hormonal.
- Enzima: biocatalizador de naturaleza proteínica, de carácter endógeno, es decir, producido por las células corporales, responsable de todos los procesos metabólicos del organismo. En cantidades mínimas produce cambios químicos, sin intervenir ella misma en la reacción.
- Ruta Metabólica: conjunto de reacciones químicas catalizadas por una enzima que va a tener un sustrato, un producto y metabolitos intermedios. En muchos casos, el producto final de una ruta metabólica es la sustancia inicial de otra ruta.
- Metabolito: es cualquier molécula utilizada o producida durante el metabolismo.
- Sustrato: metabolito de partida en una reacción metabólica.
- Producto: producto final de la vía metabólica.

3.6.2 Requisitos Funcionales de COPABI

Los requisitos funcionales son las capacidades o condiciones que el sistema debe cumplir; en este caso se plantearon los siguientes:

R1. Mostrar listado de organismos.

R2. Mostrar información de un organismo.

R3. Mostrar listado de *pathways*.

R4. Mostrar listado de *pathways* por organismo.

R5. Mostrar información de *pathway*.

R6. Mostrar listado de enzimas.

R7. Mostrar listado de enzimas por organismo.

R8. Mostrar listado de enzimas por *pathway*.

R9. Mostrar información de enzima.

R10. Mostrar listado de compuestos.

R11. Mostrar listado de compuestos por organismo.

R12. Mostrar listado de compuestos por *pathway*.

R13. Mostrar información de compuesto.

R14. Mostrar listado de genes.

R15. Mostrar listado de genes por organismo.

R16. Mostrar listado de genes por *pathway*.

R17. Mostrar información de gen.

R18. Exportar información en formato SBML.

R19. Reconstruir modelos metabólicos a escala genómica.

3.6.3 Análisis de la aplicación

Rol del sistema

Tabla 2. Descripción del actor para COPABI.

Actor	Justificación
Usuario	Es la persona que va a interactuar con COPABI para realizar la consulta de la información biológica, la descarga de la información en formato SBML Nivel 2, Versión 1 o reconstruir modelos metabólicos a escala genómica de los organismos.

3.6.4 Modelo de Casos de Uso de COPABI

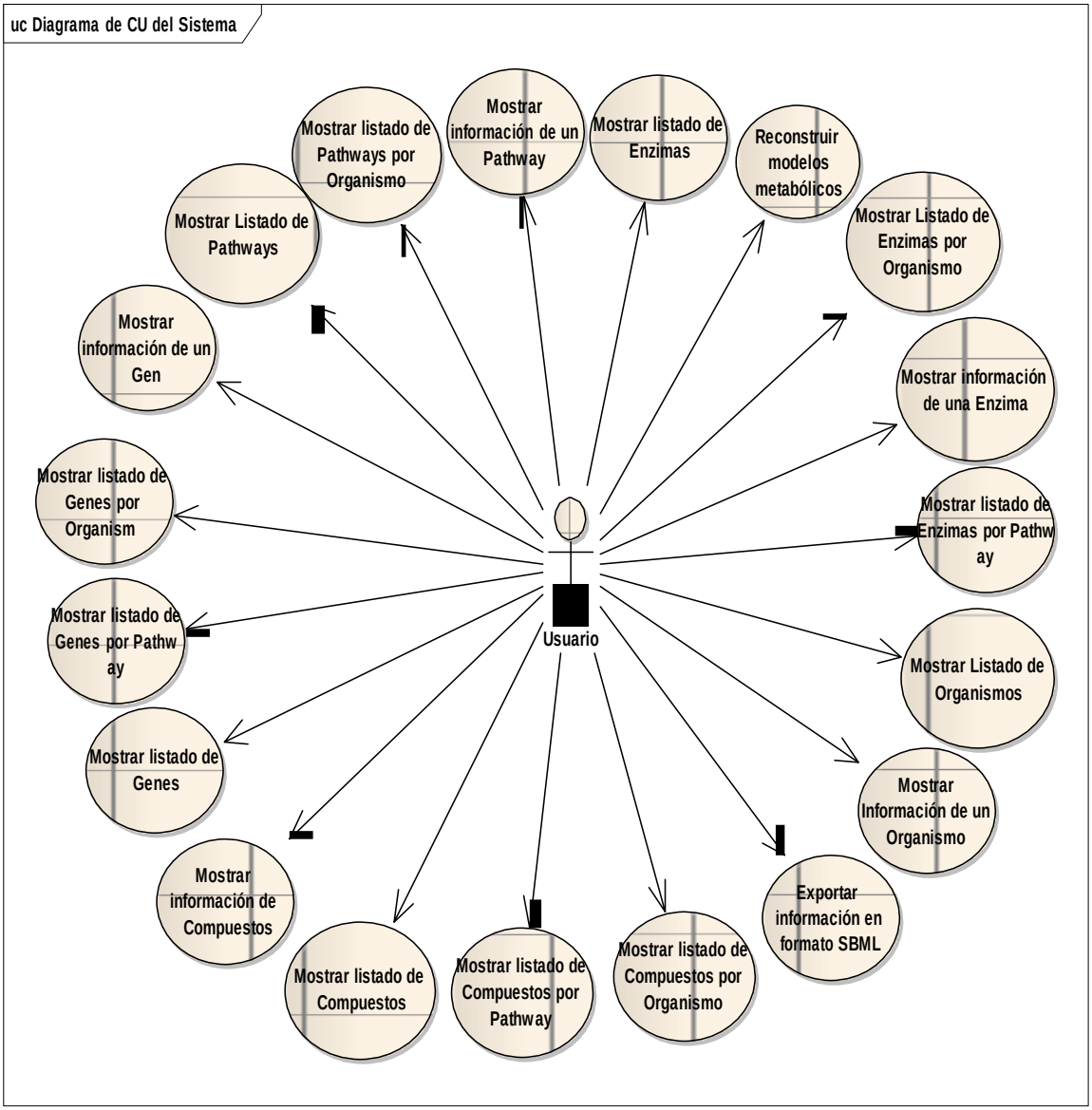

Figura 15. Diagrama de Casos de Uso de COPABI.

3.6.5 Diagramas de clases del Análisis y el Diseño

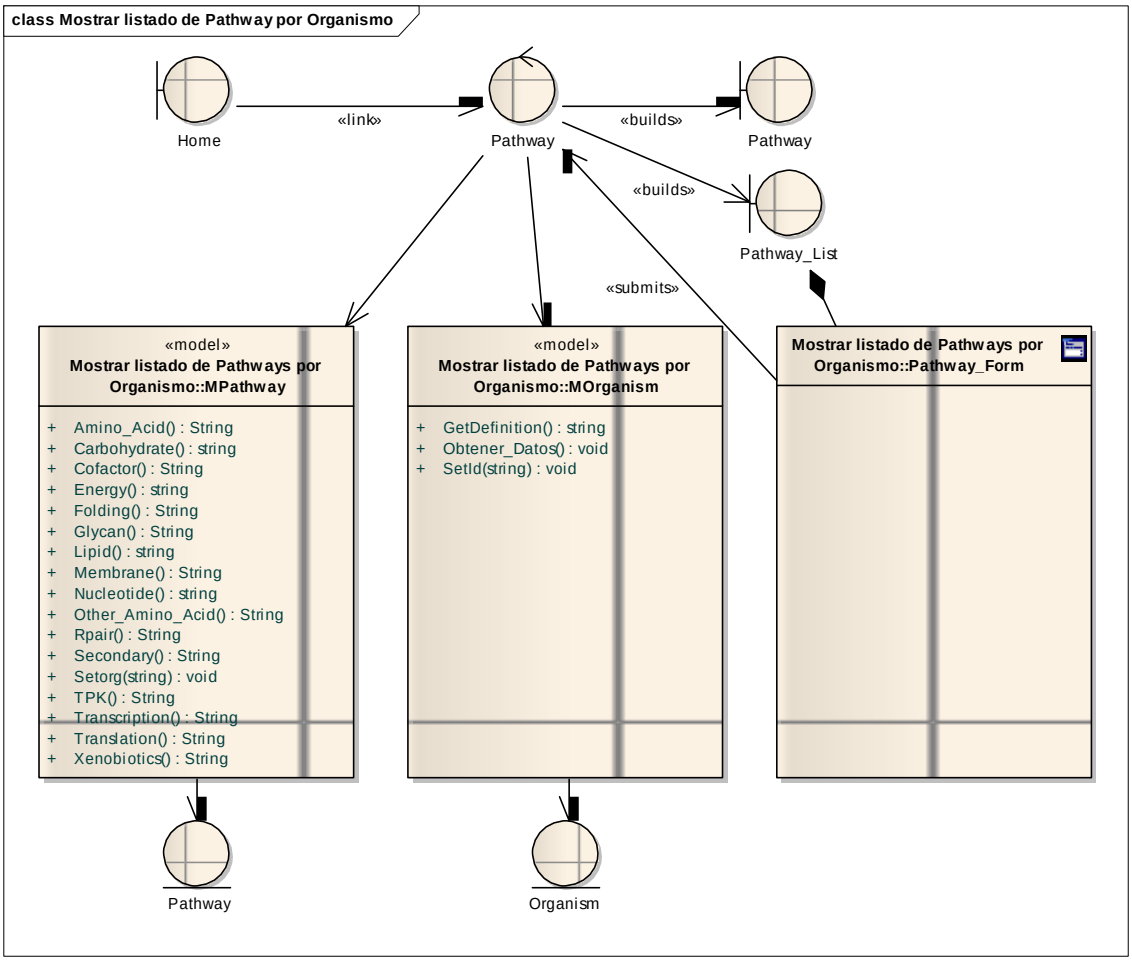

Figura 16. Diagrama de Clases del Análisis. CU_ Mostrar listado de *pathways* por organismo.

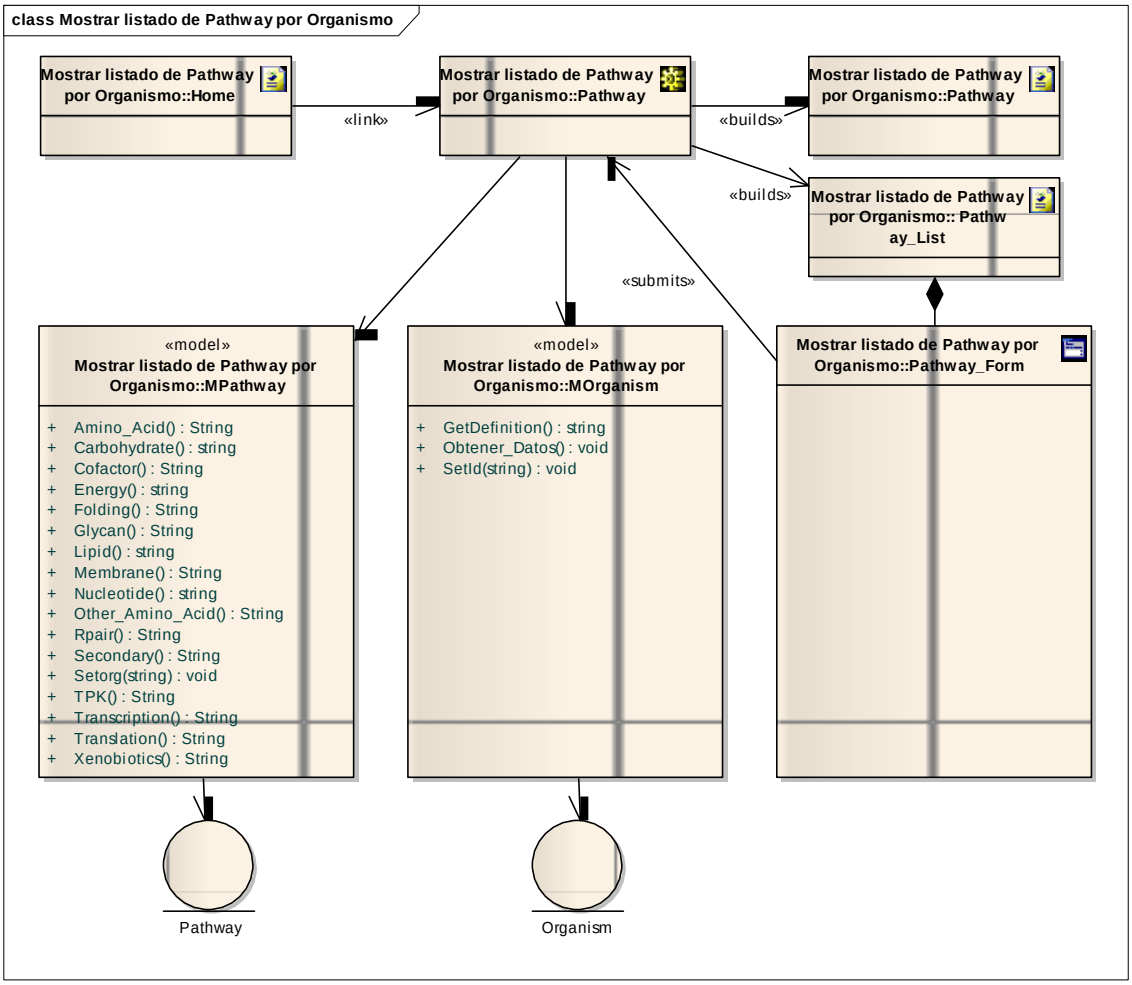

Figura 17. Diagrama de Clases del Diseño. CU_ Mostrar listado de *pathways* por organismo.

3.6.6 Modelo de componentes

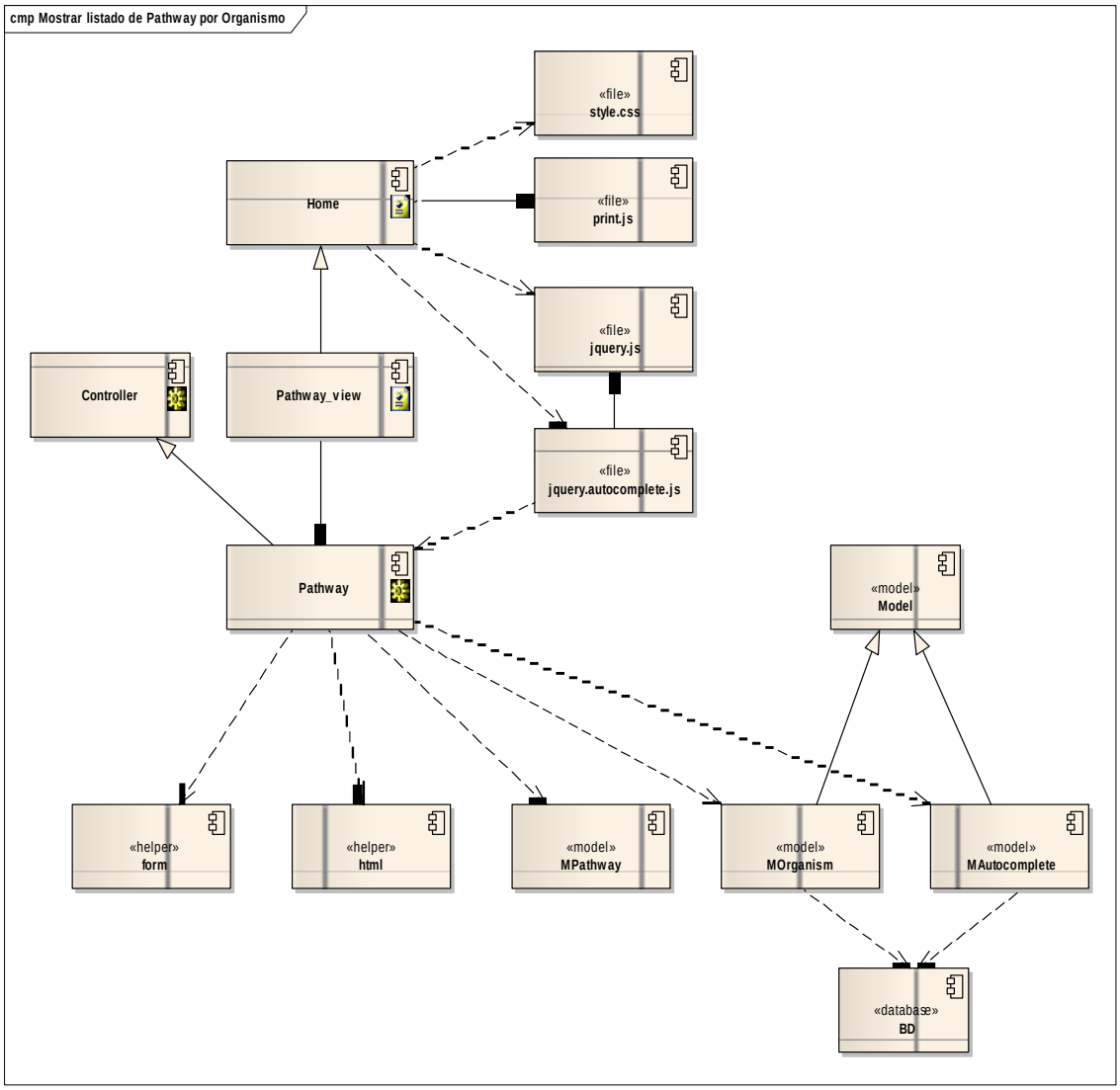

Figura 18. Modelo de componentes CU_ Mostrar listado de *pathways* por organismo.

3.6.7 Modelo de Despliegue

El Modelo de Despliegue define la arquitectura física del sistema por medio de nodos interconectados. Se utiliza para comprender las actividades de diseño e implementación debido a que la distribución del sistema permite un mejor desarrollo del diseño.

La aplicación que se propone está basada sobre una arquitectura cliente - servidor representada por cuatro nodos. El nodo “PC Cliente” contiene un navegador para

Internet, el cual recibe la información en lenguaje HTML enviado desde el servidor y se encarga de comunicarse con el nodo que contiene la aplicación web a través del protocolo HTTP. Este proceso se realiza a través de los recursos que se le muestran al usuario en la página, lo cual permite al usuario establecer un sistema de comunicación con el Servidor Web Apache. El dispositivo de salida “Impresora” se conecta al nodo “PC-Cliente” a través de:

- Puerto serie y la comunicación fluye mediante el protocolo RS-232.
- Puerto USB.
- A través de la red mediante el protocolo TCP/IP.

En el nodo “Servidor Web Apache” se atienden las solicitudes del cliente, se analizan y se les da respuesta. En este nodo están contenidos todos los procesos de información que garantizan el funcionamiento del servidor logrando cumplir con todos los requerimientos funcionales del sistema. La capa de acceso a datos se comunica con el nodo “Servidor BD Postgres” a través del protocolo TCP/IP donde se encuentra la información almacenada en la base de datos.

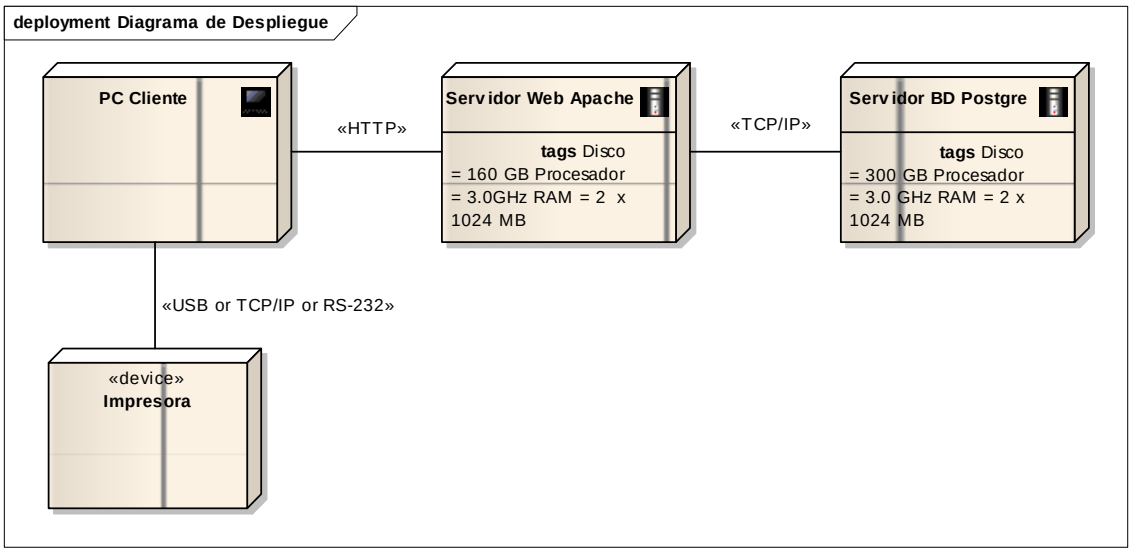

Figura 19. Diagrama de despliegue de COPABI.

3.6.8 Mapa de Navegación

La navegación de COPABI fue diseñada sobre la base del rol jugado por el actor de la aplicación, garantizando que le sean brindadas las funcionalidades que este necesita.

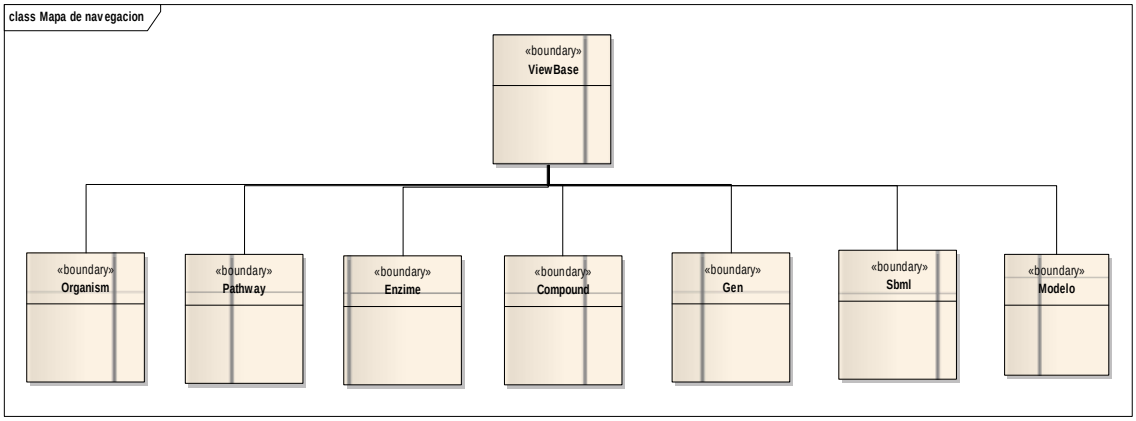

Figura 20.Mapa de navegación de COPABI.

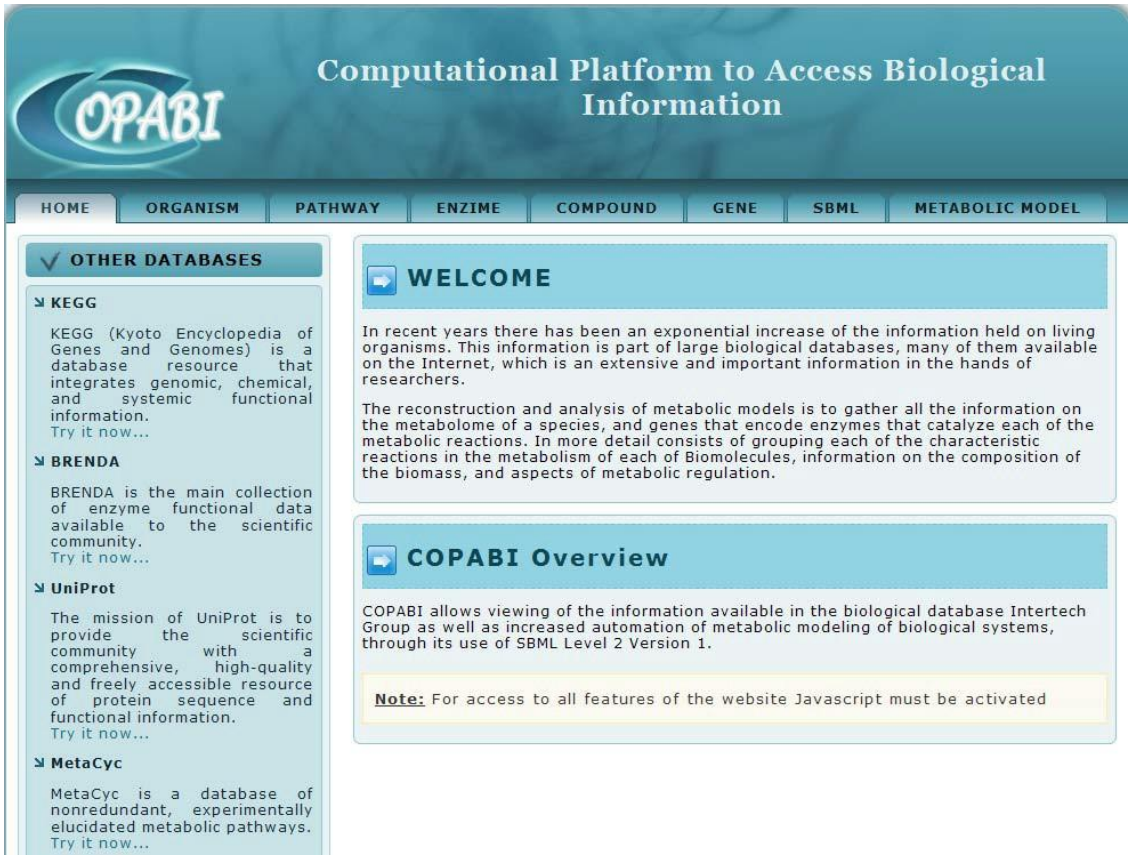

Figura 21. Interfaz principal de COPABI.

Capítulo 4

Validación de los resultados

Para la validación de los modelos metabólicos generados por COPABI, se han comparado los modelos reconstruidos automáticamente para varios organismos con aquellos reportados en la literatura y reconstruidos manualmente.

En un primer paso se analizan las propiedades generales de los modelos obtenidos (número de metabolitos, número de reacciones, porcentaje de la reversibilidad de las reacciones, cantidad de metabolitos conectados entre sí, etc.) y luego las propiedades globales de las redes descritas por el modelo metabólico.

4.1 Propiedades generales

Desde el punto de vista de la red, cada metabolito de un modelo se puede pensar como un nodo y la reacción representa los vínculos entre los metabolitos. Estos enlaces se pueden dirigir si se tiene en cuenta la reversibilidad de las reacciones.

En una primera etapa del análisis, se realiza la búsqueda en los modelos metabólicos de reacciones mal anotadas, o sea, reacciones sin sustratos o productos, (algunas de las reacciones de transporte en los archivos SBML que se tomaron de la literatura presentan este problema). Se analizan, igualmente, reacciones donde un mismo metabolito aparece como sustrato y como producto o reacciones desconectadas de la red. Esto último significa que al menos un sustrato o un producto de la reacción aparecen una única vez dentro del modelo. Reacciones con estas características son llamadas adversas y se sustraen de los modelos antes de llevar a cabo cualquier cálculo.

Cada modelo metabólico se representa como una red. Durante la comparación hemos optado por trabajar con la versión dirigida de la misma. Esto significa que los enlaces de conexión de dos metabolitos tienen una dirección desde el sustrato hasta el producto y en el caso de las reacciones reversibles los metabolitos asociados tendrían dos enlaces de direcciones opuestas que los conectan. Se han generado dos versiones del modelo metabólico para cada organismo analizado, variando el

parámetro de decisión para lograr el criterio de completitud de las rutas metabólicas (véase la sección 1.3). En un caso fue definido un 100% y un 10% en el otro, que representan dos extremos posibles. Un 100% definido en el parámetro de decisión trae consigo que se genere un modelo constituido solo por las reacciones cuyas enzimas están anotadas en el genoma de un organismo. Por otra parte, si se configura el parámetro al 10% significa que si una de cada diez reacciones en una vía está anotada en el genoma, entonces el modelo generado tendrá todas las reacciones correspondientes a dicha vía metabólica que se encuentra en el *pathway* teórico general.

Así, por cada organismo hay tres modelos de estudio. Dos reconstruidos de forma automática por COPABI y otro creado de forma manual a partir de artículos publicados. Los modelos tomados de la literatura para hacer la comparación se corresponden con los siguientes organismos: la *Synechocystissp PCC6803* [Montagud et al. (2010)], *Synechococcus elongatus sp. PCC7942* [Triana et al. (Enviado)], *Burkhoderia cenocepacia J2315* [Fang et al. (2011)], *Sphaeroides Rhodobacter* [Imam et al. (2011)], *Clostridium beijerinckii* [Milne et al. (2011)], *Mycoplasma genitalium* [Suthers et al. (2009)]*, Lactobacillus plantarum* [Teusink et al. (2006)], *Thermotoga marítima* [Zhang et al. (2009)] y *Yerisinia pestis* [Navid and Almaas, (2009)].

En la Tabla # 3 se muestran los resultados de la comparación general de los modelos metabólicos. A continuación se explica cada columna de la tabla por separado:

- # Met. → El número de diferentes metabolitos (o compuestos) que se encuentran en el modelo.
- # Reac. → El número de reacciones presentes en el modelo (después de excluir las reacciones adversas).
- % Rev. → El porcentaje de las reacciones que son reversibles.
- % Irr. → El porcentaje de las reacciones que son irreversibles.
- ASP → El promedio del camino más corto. Para cada par de metabolitos presentes en el modelo hemos utilizado el algoritmo de Dijkstra para calcular el camino más corto dentro en la red. Para todos los pares de metabolitos en los que se encontró el camino más corto, se calculó el valor promedio (los pares de metabolitos que no están conectados por ningún camino se quedaron fuera).Camino más corto entre dos pares de metabolitos significa el número de pasos que hay que recorrer para llegar desde uno hasta el otro.

- σASP → La desviación estándar para el cálculo de ASP.
- N_R → El número de pares de metabolitos de una ruta metabólica que se encuentran conectados.
- N_U → El número de pares de metabolitos de una ruta metabólica que no se encuentran conectados. Hay que señalar que la red es dirigida, con lo cual, los metabolitos que no tienen ningún vínculo apuntando en su dirección no pueden ser alcanzados por un par y, por lo tanto, son metabolitos externos que deben ser absorbidos por la célula desde el medio extracelular o se incorporan mal en el modelo.

Tabla 3. Comparación de los parámetros generales de los modelos metabólicos para diferentes organismos.

Org.	# Met.	# Reac.	% Rev.	% Irr.	ASP	σASP	N_R	N_U
syn lit	803	893	34.49	65.51	3.51	1.15	494446	150363
syn 10	707	718	37.74	62.26	3.2	0.90	355430	144419
syn 100	656	640	36.40	63.60	3.29	0.94	295093	135243
syf lit	777	847	36.01	63.99	3.55	1.19	475612	128117
syf 10	711	705	37.02	62.98	3.19	0.88	356066	149455
syf 100	655	622	35.05	64.95	3.32	0.95	292390	136635
cbe lit	732	856	27.22	72.78	3.05	0.82	409910	125914
cbe 10	752	808	40.22	59.78	3.21	0.88	412228	153276
cbe 100	693	733	38.2	61.8	3.33	0.97	335276	144973
tma lit	583	612	41.67	58.33	3.19	0.96	242290	97599
tma 10	566	614	46.09	53.91	3.06	0.83	250504	69852
tma 100	489	517	44.1	55.9	3.24	0.91	183170	55951
bcj lit	792	847	27.63	72.37	3.04	0.83	523487	103777
bcj 10	955	1018	37.03	62.97	3.25	0.88	632355	279670
bcj 100	907	948	36.29	63.71	3.32	0.92	564967	257682
mge lit	342	262	40.08	59.92	3.00	0.99	83279	33685
mge 10	268	254	48.82	51.18	2.89	0.82	54311	17513
mge 100	116	104	55.77	44.23	3.42	1.25	11543	1913
lpl lit	513	526	31.37	68.63	2.97	0.82	221786	41383
lpl 10	566	595	41.85	58.15	3.14	0.83	233640	86716
lpl 100	492	512	41.8	58.2	3.23	0.88	173827	68237
rsp lit	788	863	64.31	35.69	2.74	0.68	593663	27281
rsp 10	869	934	41.65	58.35	3.16	0.82	543333	211828
rsp 100	827	873	40.21	59.79	3.24	0.87	485209	198720
ypk lit	817	948	29.85	70.15	3.04	0.85	339398	142238
ypk 10	838	945	39.47	60.53	3.18	0.84	520075	182169
ypk 100	779	891	39.62	60.38	3.25	0.89	444404	162437

Los resultados obtenidos reflejan la similitud entre los valores de los parámetros estudiados, lo que proporciona confiabilidad al algoritmo implementado para la reconstrucción de los modelos metabólicos a escala genómica a partir de COPABI.

4.2 Conectividad de los nodos

Como se observa en la tabla #3, aunque las redes suelen tener cientos de metabolitos diferentes, dos de ellos diferentes están, como promedio, a solo tres pasos el uno del otro (véase la columna ASP de la tabla #3). Como consecuencia, toda la red puede responder rápidamente a los cambios en las concentraciones de los metabolitos o cualquier otra perturbación. Esta proximidad de los nodos en la red se conoce como comportamiento del micromundo y es consecuencia de una propiedad relacionada con la conectividad de la red denominada distribución libre de escala [Faloutsos et al. (1999)]. Esto significa que la distribución de la conectividad está dada por una ley potencial [Boccalettiet al. (2006)], en la que se tiene que el número de nodos (p) con un número de conexiones (x) sigue una ley de potencia: $P\,(x) \sim x^{-\gamma}$ donde γ es por lo general un número entre 2 y 3. De esta ley se concluye que hay muy pocos nodos con un gran número de conexiones (que son llamados *hubs*) y la mayoría de los nodos tienen muy pocas conexiones. En el estudio de la conectividad de los nodos, se implementó un algoritmo que cuenta, por cada metabolito, la cantidad de reacciones en las que aparece como un sustrato (o producto en las reacciones reversibles), así como la conectividad con las enzimas.

A continuación se representa para varios organismos analizados, los resultados de la comparación entre los dos modelos automáticos generados por COPABI y el modelo reportado en la literatura. En ellos se puede ver claramente la tendencia de la distribución siguiendo la ley de potencia y la similitud entre los modelos, demostrando la efectividad de la metodología implementada para la reconstrucción de modelos metabólicos a escala genómica.

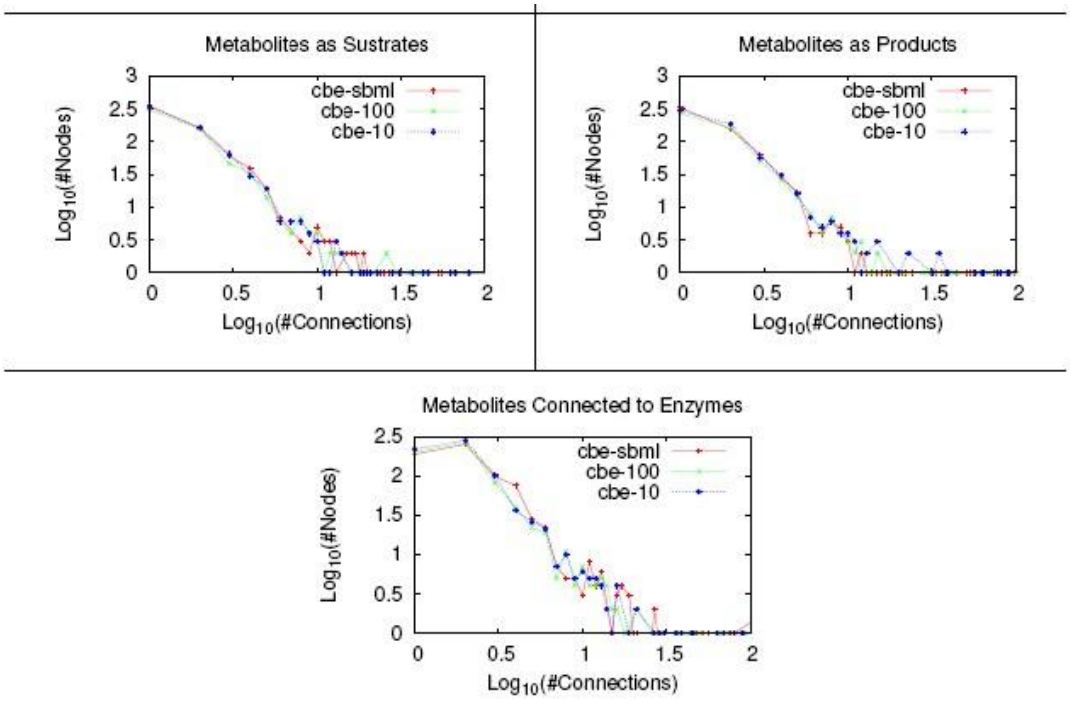

Figura 22. Distribución de la conectividad de metabolitos como sustratos y como producto y la conectividad de metabolitos con enzimas para el *Clostridium beijerinckii*.

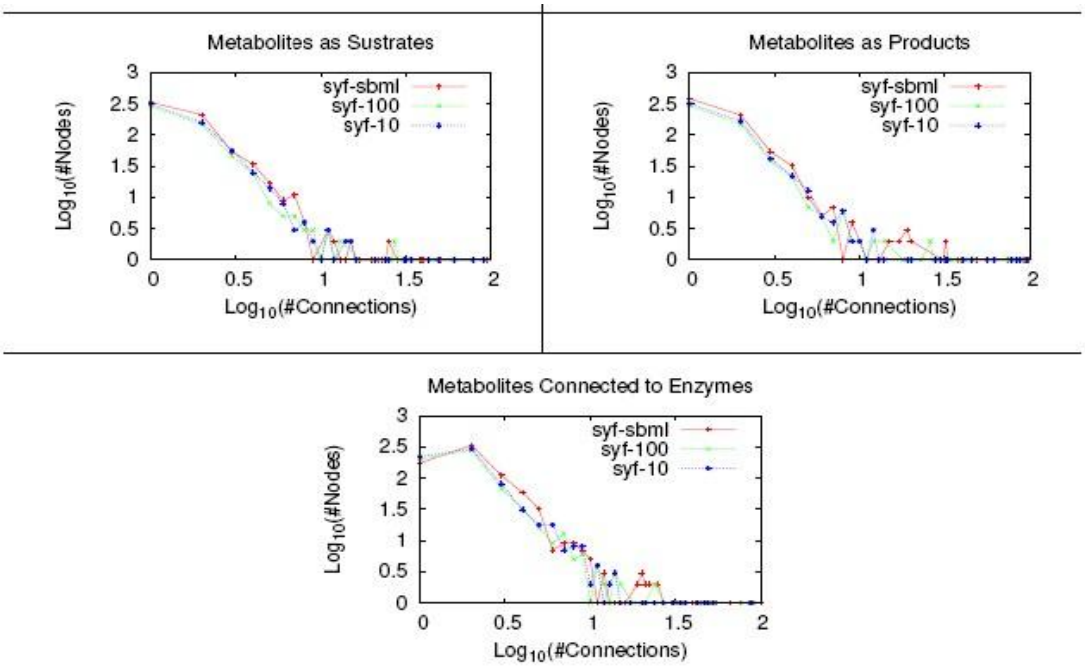

Figura 23. Distribución de la conectividad de metabolitos como sustratos y como producto y la conectividad de metabolitos con enzimas para el *Synechococcus elongatus PCC 7942*.

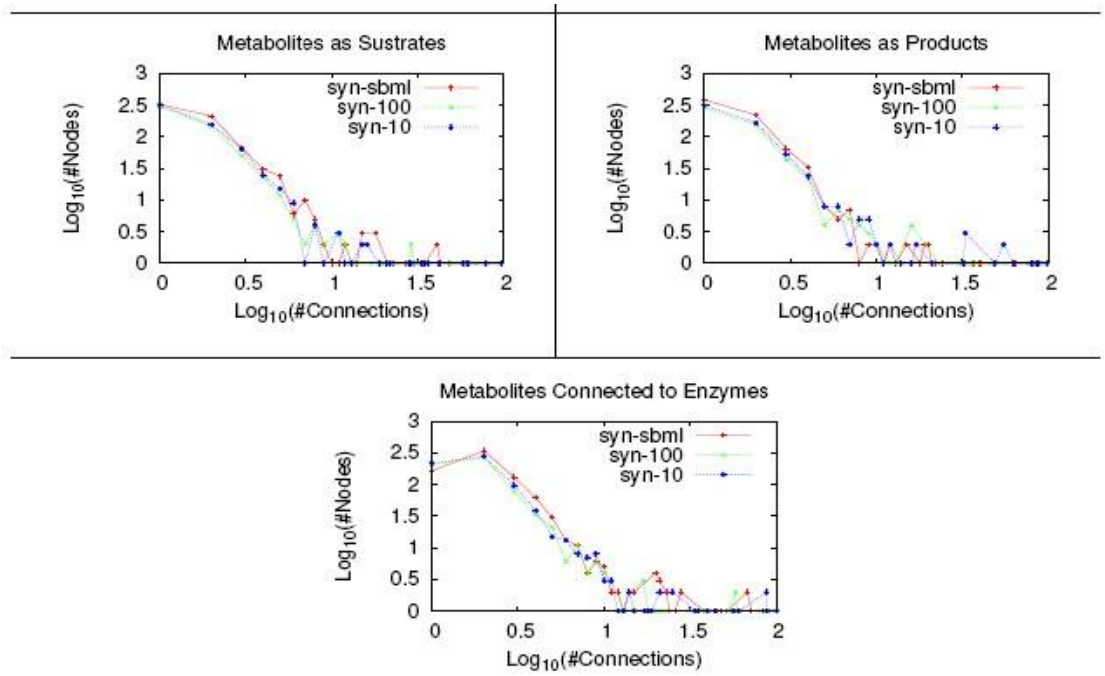

Figura 24.Distribución de la conectividad de metabolitos como sustratos y como producto y la conectividad de metabolitos con enzimas para el *Synechocystis sp. PCC 6803*.

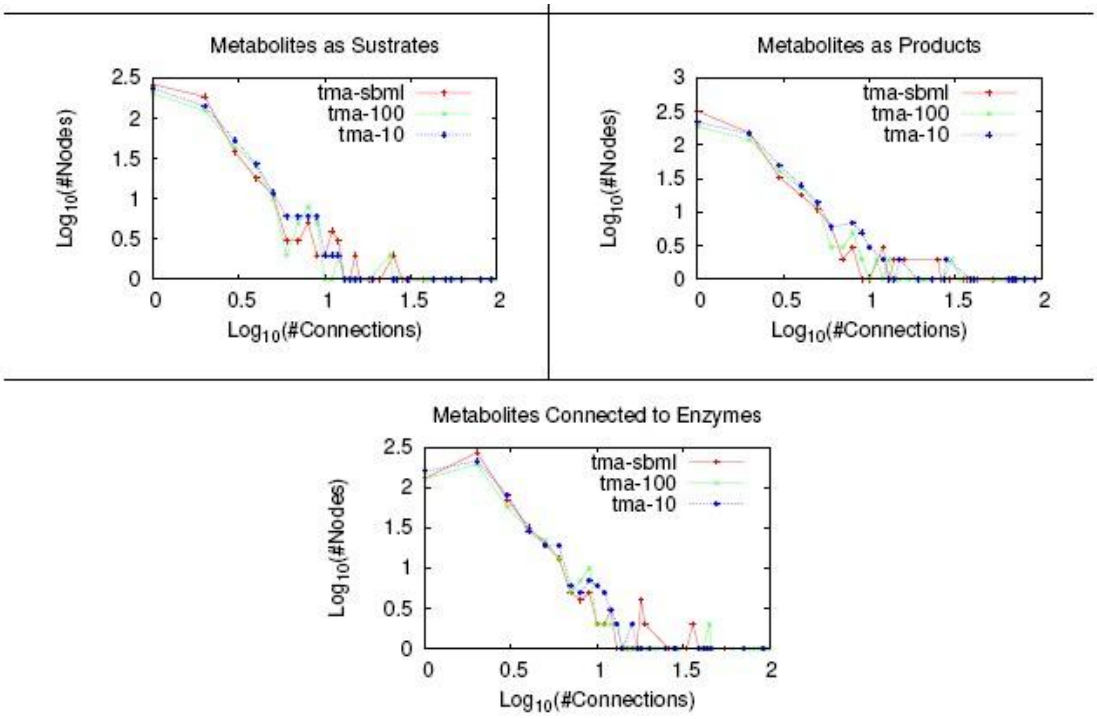

Figura 25. Distribución de la conectividad de metabolitos como sustratos y como producto y la conectividad de metabolitos con enzimas para la *Thermotoga marítima*.

Hasta aquí hemos podido apreciar que todas las redes metabólicas son muy similares cuando se estudian sus propiedades. Por lo tanto, con el fin de ser capaz de diferenciar la red metabólica de un organismo de la de otro distinto, se deben analizar los detalles de las redes, es decir, los metabolitos específicos y los *hubs* que son particulares para uno u otro organismo. En este sentido, se definió un parámetro que midiera el grado de similitud de dos redes metabólicas. Dos elementos se tuvieron en cuenta en la definición de este parámetro, en primer lugar los metabolitos presentes en cada red metabólica y el grado (número de conexiones) de cada metabolito con el resto de los metabolitos presentes en la red.

4.3 Criterio para diferenciar dos redes

El objetivo principal en este paso es definir un criterio que permita diferenciarlas redes metabólicas. En este sentido, cuanto mayor sea el valor de este parámetro, más diferentes serán las redes. Por el contrario, mientras menor sea el valor, más parecidas serán.

Dadas dos redes metabólicas, cada una tiene un conjunto de metabolitos (llamémosle conjuntos A y B). Sin embargo, entre todos los metabolitos en las dos redes hay tres conjuntos diferentes: los metabolitos particulares a la redA, los metabolitos particulares a la red B y los metabolitos comunes a las dos redes:

$$A \cup B = \underbrace{(A \cap \bar{B})}_{\text{Only in A}} \cup \underbrace{(A \cap B)}_{\text{Common}} \cup \underbrace{(\bar{A} \cap B)}_{\text{Only in B}} \quad (1)$$

Si consideramos las conexiones de los metabolitos: Cada metabolito *i* tiene n_i conexiones en total y $n_{\alpha i}$conexiones solo a los metabolitos del conjunto $(A \cap \bar{B}), n_{\beta i}$conexiones solo a los metabolitos del conjunto $(\bar{A} \cap B)$y $n_{\gamma i}$conexiones solo a los metabolitos del conjunto $A \cap B$.

A continuación definiremos el número de metabolitos y el número total de conexiones dentro de cada conjunto:

$$N_\alpha = |A \cap \bar{B}| \quad (2)$$

$$N_\beta = |B \cap \bar{A}| \quad (3)$$

$$N_\gamma = |A \cap B| \qquad (4)$$

$$N_A = \sum_{i \in A \cap B} n_i \qquad (5)$$

$$N_B = \sum_{i \in B \cap A} n_i \qquad (6)$$

$$N_C = \sum_{i \in A \cap B} n_i \qquad (7)$$

Aquí, $|C|$ indica el cardinal o número de elementos en el conjunto C.

Ahora, para cada conjunto de metabolitos, se calcula el sumatorio de la proporción que representa las conexiones del metabolito con otros metabolitos pertenecientes al conjunto respecto al total de conexiones del metabolito, ponderando este valor por el resultado de la división entre el número total de conexiones internas del conjunto y el número de metabolitos para dicho conjunto.

$$pA_i = \frac{n_{\alpha i}}{n_i}$$

$$\alpha = \frac{N_A}{N_\alpha} \sum_{i \in A \cap B} \frac{1}{n_i} pA_i$$

Y análogamente, definimos β y γ para los metabolitos en los otros dos conjuntos.

Finalmente, el criterio para diferenciar dos redes se define como:

$$crit = \frac{\alpha + \beta}{2\gamma}$$

Para una red idéntica, si α y β son cero, entonces crit = 0. Para dos redes que no tengan metabolitos en común, γ = 0 y crit = ∞.

A la hora de comparar dos modelos metabólicos se debía analizar que los metabolitos importantes en un organismo pudieran ser diferentes para otro organismo, con lo cual era relevante tener en cuenta también en la comparación los metabolitos con sus diferentes conexiones, y no solo el número de conexiones asociadas a cada metabolito. Este paso es complejo, porque los nombres de los metabolitos utilizados en los modelos metabólicos extraídos de la literatura no siguen un estándar y los autores de cada modelo eligen las diferentes abreviaturas y nombres para cada compuesto. Para algunos modelos, sin embargo, los autores también han puesto a disposición de la comunidad científica el nombre de cada compuesto utilizado en sus modelos con el identificador de KEGG que les corresponde. Para estos modelos hemos sido capaces de construir un algoritmo que traduce los nombres de los metabolitos usados en el modelo metabólico de la literatura a los mismos nombres estándar de dichos metabolitos utilizados por KEGG y por lo tanto hemos logrado comparar también las identidades de los metabolitos.

A continuación presentamos la Tabla # 4, donde se muestran los resultados de la comparación entre 6 modelos de organismos publicados en la literatura y 9 generados automáticamente por COPABI. Para generar automáticamente los modelos metabólicos se toman como parámetro de decisión los siguientes valores: el 10% en el algoritmo de completitud y se mantiene el criterio de una única reacción dentro del modelo para el algoritmo de unicidad. Nótese que hay modelos de organismos como: mge lit, lpl lit y bcj lit que no se pueden comparar, precisamente porque los autores de los modelos publicados no siguen un estándar a la hora de nombrar los compuestos. Como resultado interesante de la comparación, se puede observar que el **menor valor** en cada columna es cuando **coinciden los modelos para una misma especie.**

Tabla 4. Comparación entre los modelos generados automáticamente y los tomados de la literatura a partir del criterio definido para diferenciar las redes.

org	syn lit	syf lit	cbe lit	rsp lit	ypk lit	tma lit
mge 10	1.246	1.254	0.401	0.695	1.003	1.541
lpl 10	0.815	0.755	0.121	0.317	0.476	0.834
syn 10	**0.47**	0.527	0.183	0.248	0.626	1.065
syf 10	0.54	**0.496**	0.18	0.255	0.628	0.999
cbe 10	0.697	0.699	**0.076**	0.212	0.413	0.814

bcj 10	0.708	0.721	0.156	0.183	0.459	1.063
tma 10	0.72	0.7	0.103	0.278	0.498	**0.636**
rsp 10	0.735	0.741	0.157	**0.138**	0.549	1.103
ypk 10	0.772	0.782	0.12	0.181	**0.324**	0.882

Por lo tanto, se valida el planteamiento de que dos redes metabólicas serán más parecidas mientras menor sea el valor del criterio definido para diferenciar dichas redes.

La tabla anterior valida el hecho de que el modelo metabólico generado automáticamente para un organismo se ajusta mejor al modelo metabólico publicado en la literatura **precisamente** para ese organismo.

No obstante, se trata de una comparación aproximada, ya que como habíamos mencionado no existe una total concordancia en la identificación de los nombres de los metabolitos. Además, los modelos generados automáticamente por COPABI no han sido simulados mediante el análisis de balance de flujo, no hay distinción entre los metabolitos internos y los externos y se utilizó una versión de los modelos donde no se había definido la BM. Esto introduce diversos errores e incertidumbre en la comparación que se hace. Los modelos reportados en la literatura tienen estos elementos definidos y aparecen, desde el punto de vista de comparación del algoritmo, como metabolitos nuevos, para los que no habrá contrapartida en los modelos generados automáticamente. A pesar de esto, las características ya incluidas en los modelos generados de manera automática por la aplicación, son suficientes para comparar exitosamente con los modelos tomados de la literatura.

Como se pudo apreciar en este capítulo, se han comparado los modelos metabólicos generados por COPABI con los modelos publicados sobre el mismo organismo en la literatura, evaluándose las características generales de las redes. En todo momento, la comparación muestra que los modelos generados automáticamente son consistentes con los modelos construidos manualmente, lo que demuestra la eficiencia de los algoritmos probabilísticos implementados para la reconstrucción de modelos metabólicos a escala genómica de organismos. Los resultados anteriores se pueden consultar en la siguiente publicación [Reyes et al. (2012)].

Capítulo 5

Conclusiones

En la presente memoria tratamos la automatización del proceso de reconstrucción de modelos metabólicos a escala genómica. Nos basamos en la metodología implementada de forma manual para la reconstrucción del primer modelo metabólico desarrollado para un microorganismo fotosintético, la *Synechocystis sp. PCC6803* [Montagud et al. (2010)]. Este modelo fue obtenido por el grupo de Modelización Interdisciplinar InterTech, de la Universidad Politécnica de Valencia. La principal ventaja en la automatización de esta metodología radica en el desarrollo y análisis de algoritmos que incluyen decisiones basadas en criterios probabilísticos. Estos criterios se basan en la unicidad y completitud de las vías metabólicas. Como consecuencia de la aplicación de estos algoritmos se obtiene una aplicación web, COPABI, que permite reconstruir modelos metabólicos a escala genómica. Los resultados obtenidos durante la comparación de los modelos metabólicos publicados en la literatura con los generados por COPABI demuestran la efectividad de la metodología implementada para reconstruir dichos modelos.

A continuación se presentan las principales conclusiones originales de este trabajo:

1. Se ha realizado un estudio de la información disponible en las distintas bases de datos biológicas, con el fin de identificar la información que manejaba cada una de ellas y la forma de acceder a las mismas. En este sentido, se diseñó e implementó un SWC para acceder a KEGG a partir de su WS, disponible en KEGG API.
2. Se ha construido una base de datos que se sustenta en un modelo relacional, capaz de gestionar eficientemente la información genómica, proteómica y metabolómica obtenida a partir de la base de datos KEGG. Se usó Postgres como SGBD.
3. Para la automatización del proceso de reconstrucción de modelos metabólicos a escala genómica, se han identificado criterios probabilísticos que permiten satisfacer los criterios de completitud y unicidad de los modelos. En este

sentido están 1) la inclusión de rutas metabólicas adicionales. Su selección se fundamentó por la prevalencia de metabolitos en un mapa metabólico general, conformado por todas las reacciones metabólicas que existen en los sistemas vivos de la naturaleza y 2) la presencia repetida de una misma reacción metabólica pero relacionada a diferentes enzimas. Nuevamente, se usó un criterio probabilístico para la toma de decisión basada en la unicidad de las vías metabólicas, considerando una única reacción bioquímica.

4. Se ha implementado COPABI, una aplicación web sencilla y de fácil manejo que permite reconstruir modelos metabólicos a escala genómica. COPABI permite exportar la información haciendo uso del SBML(nivel 2, versión 1). En el caso específico de los modelos metabólicos, además del SBML, también se exportan siguiendo los requerimientos de entrada del software OptGene. Ambas formas son utilizadas en la identificación de determinados objetivos dentro de la Ingeniería Metabólica.
5. Se ha demostrado numéricamente la similitud entre dos modelos automáticos generados por COPABI (10% y 100%) y el modelo utilizado de la literatura, tomando como indicadores las características comunes de las redes (cantidad de reacciones, cantidad de metabolitos, % de reacciones reversibles e irreversibles, etc.).
6. Se han implementado algoritmos estándar para calcular el promedio de la ruta más corta entre los nodos de la red y la conectividad de los mismos. Para este último paso, se implementó un algoritmo que cuenta por cada metabolito, la cantidad de reacciones en las que aparece como un sustrato (o producto en las reacciones reversibles), así como la conectividad con las enzimas. Los resultados muestran la tendencia de la distribución siguiendo una ley de potencia y la similitud entre los modelos analizados.
7. Se ha demostrado numéricamente la similitud entre los modelos. El indicador en este caso fue el cálculo de un criterio definido para diferenciar redes metabólicas. Se realizó una comparación aproximada entre los modelos que genera la aplicación automáticamente con los modelos tomados de la literatura. Se utilizó el parámetro de decisión para el algoritmo de completitud al 10% y en el caso del algoritmo de unicidad se mantiene el criterio de una única reacción dentro del modelo. Los resultados obtenidos de la comparación entre 6 modelos de la literatura para diferentes organismos, demuestran que el menor valor en el cálculo de este criterio es cuando coinciden los modelos para una

misma especie, lo que evidencia la consistencia de los modelos reconstruidos automáticamente.

Bibliografía

A

[Aleksic et al. (2007)]. Aleksic, J., Bizzari, F., Cai, Y., Davidson, B., Mora, K., Ivakhno, S., Seshasayee, S.L., Nicholson, J., Wilson, J., Elfick, A., French, C., Kozma-Bognar, L., Ma, H. and Millar, A. 2007. Development of a novel biosensor for the detection of arsenic in drinking water. IET Synthetic Biology. Vol. 1, No. 40940, 2007, p. 87-90.

[Altman, (2004)].Altman, R.B. 2004. "Building successful biological databases", Brief. Bioinformatics 5 (1): 4-5, March 2004. PMID 15153301.

[Attwood et al. (2011)]. Attwood, T.K., Eriksson, A.G. and Bongcam-Rudloff, E. 2011. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective Bioinformatics - Trends and Methodologies, ISBN: 978-953-307-282-1doi: 10.5772/23535.

B

[Bairoch, (2000)]. Bairoch, A. 2000. The ENZYME database in 2000. Nucleic Acids Research 28, pages: 304–305. doi:10.1093/nar/28.1.304. PMC 102465. PMID 10592255.

[Bell, (2010)]. Bell, M. 2010. SOA Modeling Patterns for Service-Oriented Discovery and Analysis. Wiley & Sons. p. 390. ISBN: 0470481978.

[Benslimane et al. (2008)]. Benslimane,D., Schahram, D. and Sheth, A. 2008. Services Mashups: The New Generation of Web Applications. IEEE Internet Computing, vol. 12, no. 5. Institute of Electrical and Electronics Engineers. pp. 13–15.

[Boccaletti et al. (2006)]. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.U. 2006. Complex Networks: Structure and Dynamics. Physics Reports.Volume 424, Issues 4–5, Pages 175–308.

[Boele et al. (2012)]. Boele, J., Olivier, B.G. and Teusink, B. 2012. FAME, the Flux Analysis and Modeling Environment. BMC System Biology. 2012;6(8) doi: 10.1186/1752-0509-6-8.

[Booch et al. (2000)].Booch, G., Jacobson, I., and Rumbaugh, J. 2000. OMG Unified Modeling Language Specification. 1st Edition: March 2000. Retrieved 12 August 2008.

[Books, (2010)].Books, Llc. 2010. Biological Sequence Format: Fasta Format, Stockholm Format, Fastq Format. Publisher: General Books LLC, 2010. ISBN-10: 1158522533 ISBN-13: 9781158522538.

[Bourne, (2005)]. Bourne, P. 2005. Will a biological database be different from a biological journal. Computational Biology.PMCID: PMC1193993, doi: 10.1371/journal.pcbi.0010034.

C

[Camacho et al. (2009)]. Camacho, C., Madden, T., Coulouris, G., Ma, N., Tao, T. and Agarwala, R. 2008. BLAST Help. NCBI Help Manual Bethesda (MD): National Center for Biotechnology Information US.

[Caspi et al. (2008)]. Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang P., and Karp, P.D. 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Research 36 (suppl 1): D623-D631. doi: 10.1093/nar/gkm900.

[Castro, (2006)]. Castro, E. 2006. HTML, XHTML, and CSS: Visual QuickStart Guide, 6th Edition, Publisher: Peachpit Press. ISBN-10: 0-321-43084-0 ISBN-13: 978-0-321-43084-7.

[Crawford et al. (2005)]. Crawford, C. H.; Bate, G. P.; Cherbakov, L.; Holley, K. and Tsocanos, C. 2005. Toward an on demand service-oriented architecture. IBM Systems Journal, Volume: 44, Issue: 1, Page(s): 81 – 107, Digital Object Identifier: 10.1147/sj.441.0081.

CH

[Chang et al. (2009)]. Chang A, Scheer M, Grote A, Schomburg I. and Schomburg D. 2009. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Research37: Database issue 2009.

[Chen and Vitkup, (2006)]. Chen, L. and Vitkup, D. 2006. Predicting genes for orphan metabolic activitiesusing phylogenetic profiles. Genome Biol 2006, 7:R17.

[Christensen et al. (2009)] Christensen, E.F.C., Meredith, G. and Weerawarana,S. 2001. Web Services Description Language (WSDL) 1.1.Available from: http://www.w3.org/TR/wsdl.

D

[Dandekar et al. (2003)]. Dandekar, T., Moldenhauerc, F., Bulikd, S.,Bertrama, H. andSchusterc, S. 2003. A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. BioSystems 70, 255–270.

[Date and Darwen, (2009)]. Date, C. J. and Darwen, H. 1996. A Guide to the SQL standard: a user's guide to the standard database language SQL, 4th Edition, Addison Wesley, USA 1996, ISBN 978-0-201-96426-4.

[Deckwer et al. (2006)]. Deckwer, W. D, Jahn, D. , Hempel , D. and Zeng, A.P. 2006. Systems Biology Approaches to bioprocess Development. (2006) Engineering in Life Sciences. Volume 6, Issue 5, pages 455–469, October, 2006, DOI: 10.1002/elsc.200620153.

[Degtyarenko et al. (2008)]. Degtyarenko, K., Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. 2008. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research. January; 36(Database issue): D344–D350. Published online 2008 January. doi: 10.1093/nar/gkm791. PMCID: PMC2238832.

[Deitel and Deitel, (1999)]. Deitel, H.M and Deitel, P.F. 1999. Java: How to Program. Prentice Hall. 2nd Edition. ISBN0130106712, 9780130106711.

E

[Edwards et al. (1999)]. Edwards J.S., Ramakrishna R., Schilling C.H. and Palsson B.1999. Metabolic flux balance analysis. In Metabolic engineering. Edited by Lee SY and Pa-poutsakis ET. New York: Marcel Dekker Inc.; 1357.

[Edwards et al. (2001)]. Edwards, J.S., Ibarra, R.U., and Palsson, B. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nature Biotechnology. 19: 125-130.

[Elmasri and Navathe, (2007)]. Elmasri, R.A. and Navathe, S.B. 2007. Fundamentos de Sistemas de Bases de Datos. Addison-Wesley, 5ta Edición. ISBN 84-782-9085-0.

F

[Faloutsos et al. (1999)]. Faloutsos, M., Faloutsos, P. and Faloutsos, C. 1999. On power-law relationship of the internet topology. Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication. ACM New York, USA. Pages 251-262. ISBN:1-58113-135-6 doi: 10.1145/316188.316229.

[Fang et al. (2011)].Fang, K., Zhao, H., Sun, Ch., Lam, C. M., Chang, S., Zhang, K., Panda, G., Godinho, M., Martins dos Santos, V. and Wang, J. 2011. Exploring the metabolic network of the epidemic pathogen Burkholderia cenocepacia J2315 via genome-scale reconstruction. Systems biology 1752-0509/5/83.

[Feist et al. (2009)]. Feist A.M., Herrgard M.J., Thiele I., Reed J.L., and Palsson B. 2009. Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology 2009, 7:129143.

[Feng et al. (2012)]. Feng, X., Xu, Y., Chen, Y., and Tang, Y.J. 2012. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database BMC Syst Biol. 2012; 6: 94. doi: 10.1186/1752-0509-6-94.

[Förster et al. (2003)]. Förster, J., Famili, I.P.Fu, Palsson,B. and Nielsen, J.2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Research 13: 244-53.

G

[Garrido et al. (2010)]. Garrido, J., Triana, J., Villar,L., Jaime, R., Reyes, R., Córdova, V., Castro, J.C., Navarro, E., Montagud, A., F. de Córdoba, P., and Urchueguía, J.F. 2010. Rational Organism Network Painter: una herramienta optimizada de visualización de redes metabólicas de fácil uso. XV Convención Científica de Ingeniería y Arquitectura (CCIA 15). La Habana, Cuba, del 29 de noviembre al 3 de diciembre de 2010. Publicado en: Memorias de la Conferencia. ISBN 978-959-261-317-1.

[Garrido et al. (2011)]. Garrido,J., Villar, L., Reyes, R., Jaime, R., Triana,J., Córdova, V., Castro, J.C., Navarro, E., Montagud, A., F.de Córdoba, P., Urchueguía, J.F. andMartínez, J. 2011. HYDRA: una plataforma informática orientada al diseño, análisis y visualización de redes metabólicas. XIV Convención y Feria Internacional Informática 2011.La Habana, Cuba, del 7 al 11 de febrero de 2011. Publicado en: Programa Científico. Pág. 55. ISBN 978-959-7213-01-7.

[Green and Karp, (2004)]. Green, M.L. and Karp, P.D. 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics2004, 5:76.

[Gosling et al. (2005)]. Gosling, J., Joy, B., Steele, G.andBracha, G. 2005. The Java language specification, 3rdEdition. Addison-Wesley. ISBN0321246780.

[Gutmans et al. (2004)]. Gutmans, A., Bakken, S., and Rethans, D.2004. PHP 5 Power Programming. Publisher: Prentice Hall (Oct 27, 2004) ISBN-10: 0-13-147149-X ISBN- 13: 978-0-13-147149-8.

H

[Hallenbeck, (2002)]. Hallenbeck, P. 2002. Biological hydrogen production; fundamentals and limiting processes. International Journal of Hydrogen Energy, 27:1185-1193.

[Hamelinck et al. (2005)]. Hamelinck, C.N., Hooijdonk, G.V., and Faaij, A. 2005 Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle- and long-term. Biomass and Bioenergy, 28:384-410.

[Harte et al. (2004)]. Harte, N., Silventoinen, V., Quevillon, E., Robinson, S., Kallio, K., Fustero, X., Patel, P., Jokinen, P. and Lopez, R. 2004. Public web-based services from the European Bioinformatics Institute. Nucleic Acids Research 32 (suppl 2): W3-W9. doi: 10.1093/nar/gkh405.

[Harold, (2003)]. Harold, E.R. 2003. Effective XML. Addison-Wesley. pp. 10–19. ISBN 0-321-15040-6.

[Heinemann and Panke, (2006)]. Heinemann, M. and Panke, S. 2006. Synthetic biologyputting engineering into biology. 2006. Systems biology 22: 27902799.

[Henry et al. (2010)]. Henry, C., DeJongh, M., Best, A., Frybarger, P., Linsay, B. andStevens, R. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology. 2010;28(9):977–982. doi: 10.1038/nbt.1672.

[Horstmann, (2010)]. Horstmann, C.S. 2010. Big Java 4th Edition for Java 7 and 8, International Student Version by Publisher: Wiley Edition ISBN 978-0-470-55309-1.

[Hucka et al. (2003)]. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novre, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Scha,_ J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. and Wang, J. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. PubMed, PMID: 12611808.

I

[Ibarra et al. (2002)]. Ibarra, R.U., Edwards, J.S. andPalsson, B.2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420 | doi: 10.1038/nature01149.,186-189.

[Imam et al. (2011)].Imam, S., Yilmaz, S., Sohmen, U., Gorzalski, A.S., Reed, J.L., Noguera, D.R. and Donohue, T.J. 2011. iRsp1095: A genome-scale reconstruction of the Rhodobacter sphaeroides metabolic network. BMC Systems Biology: 116.

J

[Jaime et al. (2010)]. Jaime, R., Garrido, J., Reyes, R., Villar, L., Triana, J., Córdova, V., Castro, J.C., Navarro, E., Montagud, A., F. de Córdoba, P., Urchueguía, J.F., Martínez, J. and Hernández,Z. 2011. Nueva herramienta para el análisis del balance de flujo de las rutas metabólicas y su integración en Ron Painter. XIV Convención y Feria Internacional Informática 2011.La Habana, Cuba, del 7 al 11 de febrero de 2011. Publicado en: Programa Científico. Pág. 153. ISBN 978-959-7213-01-7.

K

[Kanehisa et al. (2008)]. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi ,Y. 2008. KEGG for linking, genomes to life and the environment. Nucleic Acids Research 36:480-484.

[Kanehisa and Goto, (2010)]. Kanehisa, M. and Goto, S. 2010. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 27-30.

[Karp et al. (2002)]. Karp, P.D., Paley, S. and Romero, P. 2002. The Pathway Tools software. Bioinformatics 2002, 18 Suppl 1:S225-32.

[Karp et al. (2005)]. Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. and Lopez-Bigas, N. 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research 2005, 19:6083-89.

[Kelder et al. (2009)]. Kelder,T.,P. van Iersel, M.,Hanspers, K., Kutmon, M., Conklin, B.R., Evelo, C.R. and Pico, A.R. 2012. WikiPathways: building research communities on biological pathways. Nucleic Acids Research. 2012 January; 40(D1): D1301–D1307.doi: 10.1093/nar/gkr1074.

[Kharchenko et al. (2004)]. Kharchenko, P., Vitkup, D. and Church, G.M. 2004. Filling gaps in a metabolicnetwork using expression information. Bioinformatics 2004, 20Suppl 1:I178-I185.

[Kharchenko et al. (2006)]. Kharchenko, P., Chen, L., Freund, Y., Vitkup, D. and Church, G.M. 2006. Identifyingmetabolic enzymes with multiple types of association evidence.BMC Bioinformatics 2006, 7:177.

[Krieger et al. (2004)]. Krieger, C.J., Zhang, P., Mueller, L.A., et al. 2004 MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Research, 32:D438-42.

[Korth et al. (2006)]. Korth, H.F, Silberschatz, A. and Sudarshan, S. 2006.Fundamentos de Bases de Datos. McGraw-Hill, 5ta edición. ISBN 84-481-4644-1.

[Kumar et al. (2007)]. Kumar, V.S, Dasika, M.S and Maranas, C.D. 2007.Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 2007, 8:212 doi:10.1186/1471-2105-8-212.

L

[Latendresse et al. (2012)]. Latendresse, M., Krummenacker, M., Trupp, M., and Karp, P.D. 2012. Construction and completion of flux balance models from pathway databases. Bioinformatics Vol. 28 no. 3 2012, pages 388–396. doi:10.1093/bioinformatics/btr681).

[Liao et al. (2011)]. Liao, Y.C., Chen, J.C., Tsai, M.H., Tang, Y.H., Chen, F.C. and Hsiung, C.A. 2011. MrBac: a web server for draft metabolic network reconstructions for bacteria. Bioeng Bugs. 2011 Sep-Oct; 2(5):284-7.

[Liao et al. (2012)]. Liao, Y.C., Tsai, M.H., Chen, F.C. and Hsiung, C.A.2012. GEMSiRV: a software platform for GEnome-scale metabolic model simulation, reconstruction and visualization. Bioinformatics. 2012; 28(13):1752-8

[Loogeret al. (2003)]. Looger, LL., Dwyer, M.A., Smith, J.J., and Hellinga, H.W. 2003. Computational design of receptor and sensor proteins with novel functions. Nature 423, 185-190. doi:10.1038/nature01556.

[López et al. (2007)]. López, M., Ruiz, G. and Vega, M. 2007. Informe de vigilancia tecnológica. Edición: Cintia Refojo, Genoma España.Referencia: GEN-ES07003, ISBN: 84-609-9762-6.

[Lovley, (2003)]. Lovley, D.R. 2003. Cleaning up with genomics: applying molecular biology to bioremediation. Nature Reviews Microbiology 1, 35-44, doi:10.1038/nrmicro731.

M

[Mesarovic, (1968)]. Mesarovic, M. D. 1968. Systems Theory andBiology-View of a Theoretician. SystemTheory and Biology, ed. M. D. Mesarovic´, pp. 59-87. Springer-Verlag.

[Miller et al. (2009)]. Miller, H., Norton, C.N. and Sarkar, I.N. 2009. GenBank and PubMed: How connected are they?. BMC Research Notes 2009, 2:101 doi:10.1186/1756-0500-2-101.

[Milne et al. (2011)]. Milne, C.B., Eddy, J.A., Raju, R., Ardekani, S., Kim, P.J., Senger, R.S., Jin, Y.S., Blaschek, H.P. and Price, N.D. 2011. Metabolic network reconstruction and genome scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. BMC Systems Biology, 5:130, doi:10.1186/1752-0509-5-130.

[Montagud et al. (2010)]. Montagud, A., Navarro, E., F. de Córdoba, P., Urchueguía, J.F. and Raosaheb, K. 2010. Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. Systems biology 1752-0509/4/156.

[Montagud et al. (En preparación)]. Montagud, A., .F. de Córdoba, P., and Urchueguía, J.F. Synechocystis sp. PCC 6803 metabolic models study for the enhanced production of biofuels. Manuscript in preparation.

[Myatt, (2008)]. Myatt, A. 2008. Pro Netbeans IDE 6 Rich Client Platform Edition 1st Edition. Apress. pp. 491. ISBN 1-59059-895-4.

N

[Navarro et al. (2009)]. Navarro, E., Montagud, A., F. de Córdoba, P., and Urchueguía, J.F. 2009. Metabolic flux analysis of the hydrogen production potential in Synechocystis sp. PCC6803. International Journal of Hydrogen Energy 2009, 34:8828-8838.

[Navid and Almaas, (2009)]. Navid, A. and Almaas, E. 2009. Genome-scale reconstruction of the metabolic network in Yersinia pestis, strain 91001. Molecular Systems Biology, 5, 368375.

[NCBI, (2012)]. Entrez, The life Sciences Search Engine. [http://www.ncbi.nlm.nih.gov/].

[Notebaart et al. (2006)]. Notebaart, R.A., van Enckevort, F., Francke, Ch., Siezen, R.J and Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks, BMC Bioinformatics 2006, 7:296 doi:10.1186/1471-2105-7-296

O

[Oberhardt et al. (2009)]. Oberhardt, M.A., Palsson, B. and Papin, J.A. 2009. Applications of genome-scale metabolic reconstructions. Molecular Systems Biology 5: 320, doi:10.1038/msb.2009.77.

[Osterman and Overbeek, (2003)]. Osterman, A. and Overbeek, R. 2003. Missing genes in metabolic pathways:a comparative genomics approach. Curr Opin Chem Biol 2003,7:238-251.

P

[Pacheco et al. (2011)]. Pacheco, Y., Reyes R., Triana, J. 2012. Servicio Web Cliente Orientado a la obtención de la información biológica disponible en la Base de Datos KEGG. Media: Paperback Book, 88 pages. Editorial: Editorial Académica Española. Publication Date: Apr. 6th, 2012. ISBN-10: 3848471051. ISBN-13: 9783848471058.

[Patil et al. (2004)]. Patil, K.R., Akesson, M. and Nielsen, J. 2004. Use of genome-scale microbial models for metabolic engineering. 2004. Current Opinion in Biotechnology, 15:64-9.

[Patil et al. (2005)]. Patil, K.R., Rocha, I., Förster, J. and Nielsen, J. 2005 Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics, 6:308.

[Pedrola et al. (En preparación)]. Pedrola, L., Montagud, A., Belda, E., Martínez-Blanch, J.F., Navarro, E., Porcar, M., Peretó, J., Moya, A., Ramón, D., and Urchueguía, J.F. Media optimization for ethanol production with designed yeast mutants. Manuscript in preparation.

[Peri et al. (2003)]. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., T.K.B. Gandhi, Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., et al. 2003. Development of Human Protein Reference Database as an Initial Platform for Approaching. Genome Research 13: 2363-2371. doi: 10.1101/gr.1680803.

[Pevsner, (2009)]. Pevsner, J. 2009. Front matter, in Bioinformatics and Functional Genomics, Second Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470451496.fmatter. Print ISBN: 9780470085851. Online ISBN: 9780470451496.

[Pinney et al. (2005)]. Pinney, J.W., Shirley, M.W., McConkey, G.A. and Westhead, D.R. 2005. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Nucleic Acids Research, 2005, Vol. 33, No. 4 1399–1409. doi:10.1093/nar/gki285

[Pinto et al. (2011)]. Pinto, F., Elburg, K.A., Pacheco, C., Lopo, M., Noirel, J., Montagud, A., Urchueguía, J.F., Wright, P., and Tamagnini, P. 2011 Construction of a chassis for hydrogen production: physiological and molecular characterization of a Synechocystis sp. PCC 6803 mutant lacking a functional bidirectional hydrogenase. Microbiology (Reading, England) 2011, 158:448-464.

[Pitkänen et al. (2008)]. Pitkänen, E., Akerlund, A., Rantanen, A., Jouhten, P. and Ukkonen, E. 2008. ReMatch: a web-based tool to construct, store and share stoichiometric metabolic models with carbon maps for metabolic flux analysis. J Integr Bioinform. 2008; 5(2). doi: 10.2390/biecoll-jib-2008-102.

[Pons et al. (2005)]. Pons, O., Marín, N., Medina, J.M., Acid, S. and Vila, M.A. 2005. Introducción a las Bases de Datos: El modelo relacional. Editorial Paraninfo, ISBN 8497323963.

R

[Reyes et al. (2011)]. Reyes, R., Jaime, R., Garrido, J., Triana, J., Villar, L., Córdova, V., Castro, J.C., Navarro, E., Montagud, A., F.de Córdoba, P., Urchueguía, J.F. and Martínez. J. 2011. Base de datos biológica orientada a la automatización del proceso de construcción de modelos a escala genómica. Resultados en la Synechocystis SP PCC6803. XIV Convención y Feria Internacional Informática 2011. La Habana, Cuba, del 7 al 11 de febrero de 2011.Publicado en: Programa Científico. Pág. 215. ISBN 978-959-7213-01-7.

[Reyes et al. (2011)]. Reyes, R., Garrido, J., Jaime, R., Córdova, V.,Triana, J., Villar, L., Castro, J.C., F.de Córdoba, P., Urchueguía, J.F., Navarro, E. and

Montagud, A. 2011. Desarrollo de una plataforma computacional para el modelado metabólicode microorganismos. Nereis. Revista Iberoamericana de Métodos, Modelización y Simulación Interdisciplinar.Universidad Católica de Valencia "San Vicente Mártir". 3: 25-31. ISSN 1888-8550.

[Reyes et al. (2012)]. Reyes R., Gamermann D., Montagud A., Fuentes D., Triana J., Fernández de Córdoba P., Urchueguía J. 2012. Automation on the generation of genome scale metabolic models. Journal of Computational Biology, December 2012, 19(12): 1295-1306. doi:10.1089/cmb.2012.0183.

[Reyes et al. (En preparación)]. Reyes R., Pacheco, Y., Triana J., Gamermann D., Montagud A., Fernández de Córdoba P., Urchueguía J. Integrated database for metabolic models reconstruction using COPABI. Manuscript in preparation.

[Reyes et al. (2010)]. R. Reyes, R. Jaime, J. Garrido, J. Triana, L. Villar, V. Córdova, J.C. Castro, E. Navarro, A. Montagud, P. Fernández de Córdoba, J.F. Urchueguía y J. Martínez. 2010. Diseño de bases de datos biológicas, un paso hacia la automatización del proceso de construcción de modelos a escala genómica. XV Convención Científica de Ingeniería y Arquitectura (CCIA 15).La Habana, Cuba, del 29 de noviembre al 3 de diciembre de 2010.Publicado en: Memorias de la Conferencia. ISBN 978-959-261-317-1.

[Ro et al. (2006)]. Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J. Chang, M.C., Withers, S.T., Shiba, Y., Sarpong, R. and Keasling, J.D. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature. 440.

S

[Schilling et al. (1999)]. Schilling, C.H., Schuster, S., Palsson, B. and Heinrich, R. 1999. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnology Progress 1999;15: 296303.

[Schlossnagle, (2007)]. Schlossnagle, G. 2007. Advanced Php Programming: Developing Large ScaleWeb Applications With Php 5. Publisher: Sams ISBN-10: 0672329239 ISBN13: 9780672329234.

[Segre et al. (2002)]. Segre, D., Vitkup, D. and Church, G. 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A, 99:15112- 7.

[Smith, (2010)].Smith, G. 2010. PostgreSQL 9.0 High Performance Publisher: Packt Publishing ISBN 184951030X. ISBN13: 9781849510301.

[Snell et al. (2001)]. Snell, J., Doug Tidwell and Pavel Kulchenko. 2001. Programming Web Services with SOAP. O'Reilly Media 2001. Chapter 3.

[Snoep et al. (2006)]. Snoep, J. L., Bruggeman, F., Olivier, B.G. and Westerho, H.V. 2006. Towards building the silicon cell: A modular approach. BioSystems 83: 207216.

[Sparx Systems, (2011)].Sparx Systems Argentina - SOLUS S.A. 2012. Enterprise Architect 7.0 - Modelado avanzado con UML 2.1. Obtenido de http://www.sparxsystems.com.ar/products/ea.html

[Stefanov, (2010)]. Stefanov, S. 2010. JavaScript Patterns. Publisher: O'Reilly Media; 1st Edition, ISBN-10: 0596806752 ISBN-13: 9780596806750.

[Sun and Zeng, (2004)]. Sun, J. and Zeng, A.P. 2004. IdentiCS – Identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence BMC Bioinformatics, BMC Bioinformatics 2004, 5:112 doi:10.1186/1471-2105-5-112).

[Suthers et al. (2009)]. Suthers, P.F., Dasika, M.S., Kumar, V.S., Denisov, G., Glass, J.I., et al. 2009. A Genome-Scale Metabolic Reconstruction of Mycoplasma genitalium, iPS189. PLoS Comput Biol 5(2): e1000285. doi:10.1371/journal.pcbi.1000285.

T

[Teusink et al. (2006)]. Teusink, B., Wiersma, A., Molenaar, D., Francke, C., de Vos, W.M., Siezen, R.J. and Smid, E.J.. 2006. Analysis of Growth of Lactobacillus plantarum WCFS1 on a Complex Medium Using a Genome-scale Metabolic Model. The Journalof Biological Chemistry Vol. 281, No. 52, pp. 4004140048.

[The UniProt Consortium, (2007)]. The Universal Protein Resource (UniProt). Nucleic Acids Research, 35:D193-7.

[Thiele and Palsson, (2010)]. Thiele I. and Palsson B.Ø. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc5, 93–121.

[Triana et al. (2010)]. Triana, J., Córdova, V., Jaime, R., Reyes, R., Garrido, J., Villar, L., Márquez, F., Castro, J.C., Navarro, E., Montagud, A., P. F. de Córdoba, P., andUrchueguía, J.F. 2010. Modelo metabólico de una cianobacteria, una fuente de energía a partir de la luz. I Congreso Internacional de Ingeniería Química, Biotecnológica y Alimentaria (CIIQBA 2010).La Habana, Cuba, del 29 de noviembre al 3 de diciembre de 2010. Publicado en: Memorias de la Conferencia. ISBN 978-959-261-317-1.

[Triana et al.(Enviado)]. Triana, J., Montagud, A., Gamermann, D., Siurana, M., Torres, J., Tena, J., Urchueguía J.F. and Fernández de Córdoba, P. Reconstruction procedure to generate and evaluate genome-scale

metabolic network models, use of Synechococcus elongatus PCC7942 as a case study. Submitted at Methods in Molecular Biology.

[Tong, (2008)]. Tong, K. 2008. Developing Web Services with Apache Axis2 (3rd edition). Publisher: TipTec Development. ISBN-10: 9993792918 ISBN-13: 978-9993792918.

U

[Ullman et al. (2001)]. Ullman, J.D., García-Molina, H. andWidom, J. 2001. Database Systems: The Complete Book. 1st Prentice Hall PTR Upper Saddle River, NJ, USA 2001 ISBN:0130319953.

[Upton, (2007)]. Upton, D. 2007. CodeIgniter for Rapid PHP Application Development: Improve your PHP coding productivity with the free compact open-source MVC CodeIgniter framework!. Publisher: Packt Publishing Ltd. ISBN: 978-1-847191-74-8.

V

[Varma and Palsson, (1993)]. Varma, A. and Palsson, B. 1993. Metabolic capabilities of Escherichia coli: II. Optimal growth patterns. J Theor Biol 1993, 165:503522.

[Varma and Palsson, (1994)]. Varma, A. and Palsson, B. 1994. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product excretion in wildtype Escherichia coli W3110. Applied Environmental Microbiology 1994;60:372431.

W

[Weise et al. (2006)]. Weise, S., Grosse, I., Klukas, C., Koschtzki, D., Scholz, U., Schreiber, F. and Junker, B.H. Meta-All: a system for managing metabolic pathway information. BMC Bioinformatics 2006, 7:465.

[Weiss, (2007)]. Weiss, R., 2007. From bacteria to stem cells, proceedings of the annual conference in functional genomics, Goteborg, Sweden.

[Wittig, (2009)]. Wittig, U. 2009. SABIO-RK: Curated Kinetic Data of Biochemical Reactions. Nature Proceedings: doi:10.1038/npre.2009.3085.1.

[Wu et al. (2006)]. Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N. and Suzek, B. 2006. The Universal Protein Resource (UniProt): an

expanding universe of protein information. Nucleic Acids Research. 2006 January 1; 34(Database issue): D187–D191. doi: 10.1093/nar/gkj161.

Y

[Yeager and McGrath, (1996)]. Yeager, N.J., and McGrath, R.E. 1996. Web Server Technology: The Advanced Guide for World Wide Web Information Providers. Publisher Morgan Kaufmann. ISBN 155860376X, 9781558603769

Z

[Zhang et al. (2009)]. Zhang, Y., Thiele, I., Weekes, D., Li, Z., Jaroszewski, L., Ginalski, K., Deacon, A.M., Wooley, J., Lesley, S.A., Wilson, I.A., Palsson, B., Osterman, A. and Godzik, A. 2009. Three-dimensional Structural View of the Central Metabolic Network of Thermotoga maritima. Science.18; 325(5947): 15441549. doi:10.1126/science.1174671.

Apéndices

Diagrama de secuencia para el CU: Realizar Descarga (Realizar descarga completa).

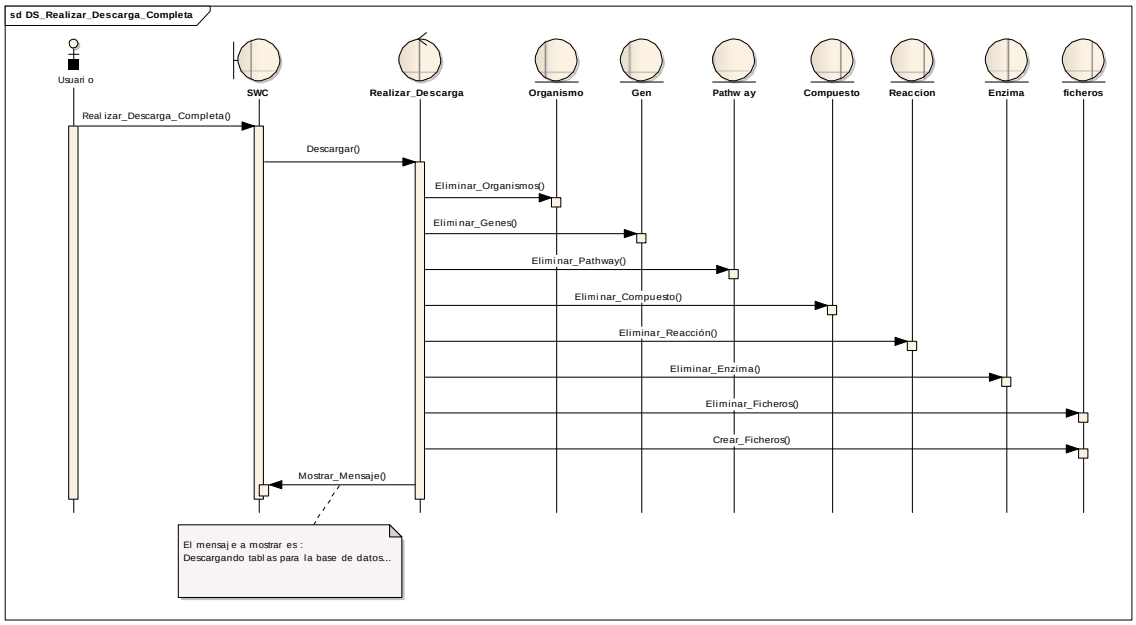

Diagrama de secuencia para el CU: Realizar Descarga (Continuar descarga).

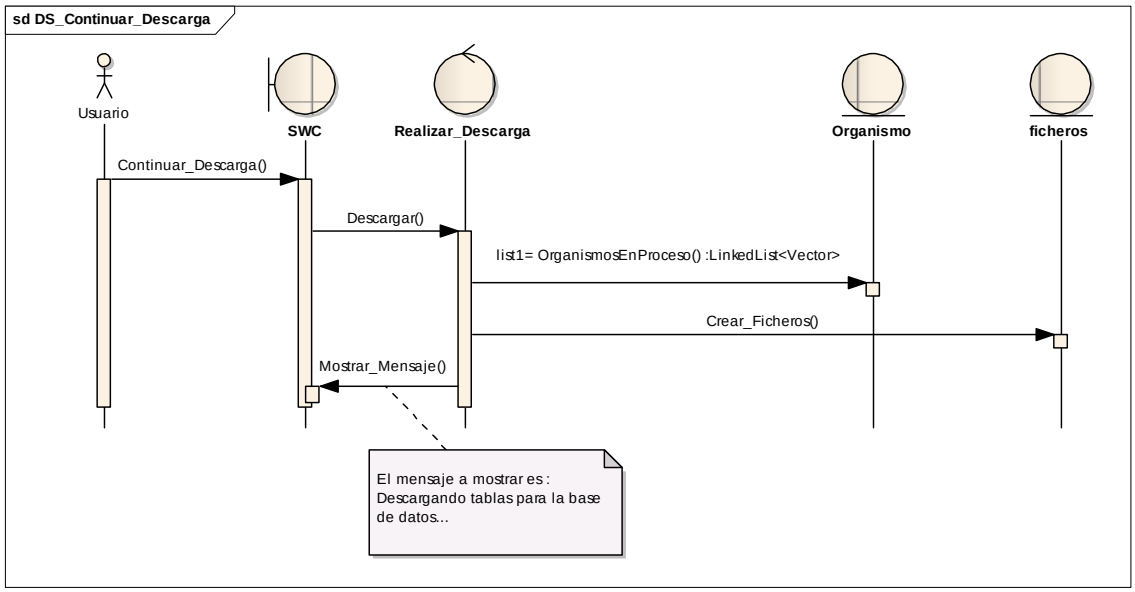

Diagrama de secuencia para el CU: Configurar conexión a Internet.

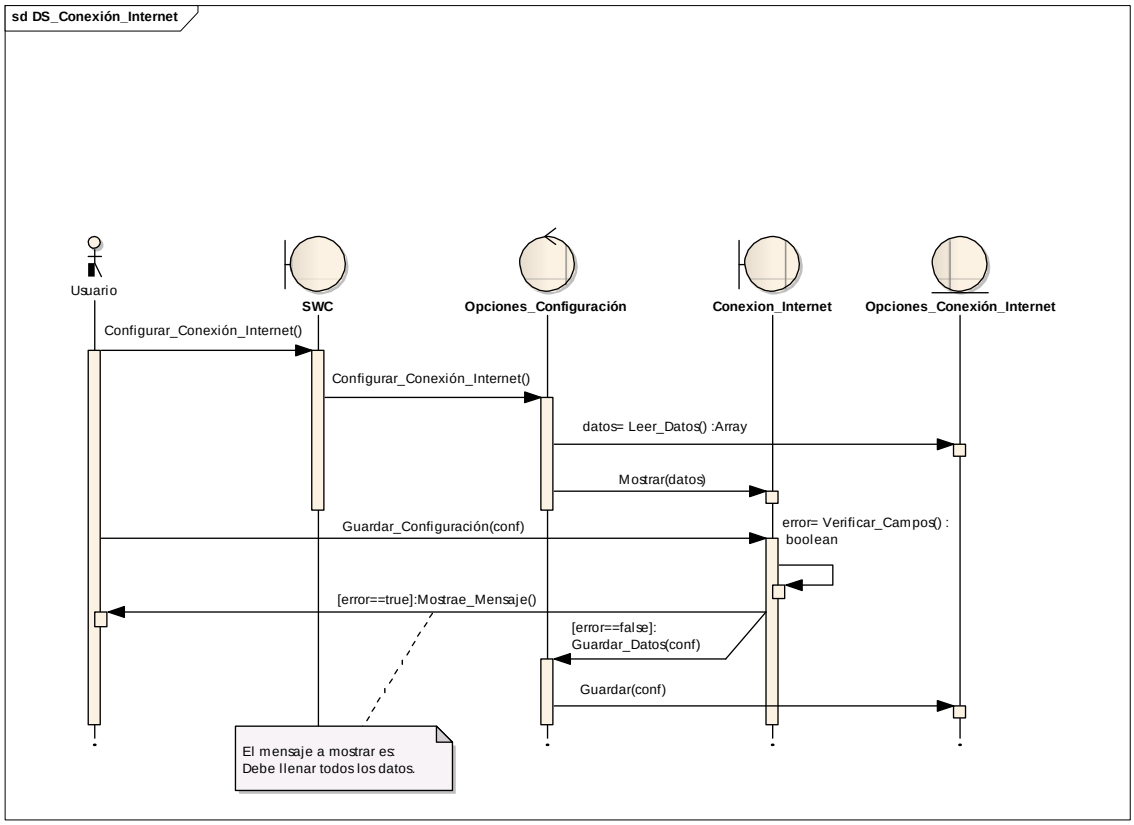

Diagrama de secuencia para el CU: Mostrar listado de pathways por organismo.