



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

IDENTIFICACIÓN DE PATRONES DE ORGANIZACIÓN DE REACCIÓN QUE
AGRUPAN NATURALMENTE TRANSFORMACIONES ENZIMÁTICAS

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias

PRESENTA:

CARLOS DANIEL VÁZQUEZ HERNÁNDEZ

TUTOR PRINCIPAL

ROSA MARÍA GUTIÉRREZ RÍOS

[Instituto de Biotecnología, UNAM campus Morelos](#)

MIEMBROS DEL COMITÉ TUTOR

Arturo Carlos II Becerra Bracho

[Facultad de Ciencias, UNAM](#)

Lorenzo Patrick Segovia Forcella

[Instituto de Biotecnología, UNAM campus Morelos](#)

Cuernavaca, Morelos. Abril, 2018



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

El presente trabajo se realizó bajo la asesoría de la Dra. Rosa María Gutiérrez Ríos en el laboratorio adscrito al Departamento de Microbiología Molecular del Instituto de Biotecnología de la Universidad Nacional Autónoma de México, campus Morelos.

Este proyecto se realizó con recursos del programa de apoyo de la Dirección General de Asuntos del Personal Académico (DGAPA) bajo el Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) con número de proyecto IN204515.

Durante mis estudios de doctorado, fui becado por el Consejo Nacional de Ciencia y Tecnología (CONACyT) bajo el Programa de Maestría y Doctorado en Ciencias Bioquímicas.

Se agradece a Ricardo Ciria Merce el apoyo en infraestructura de cómputo.

Agradecimientos

A mis padres, Carlos Vázquez y Olivia Hernández, y a mi hermana, Olivia, por su paciencia y apoyo incondicional.

A mi tutora, Rosa María Gutiérrez, por su empeño y su ejemplo de cada día.

A Enrique Merino, por ser un inigualable jefe de grupo y ejemplo de generosidad, trabajo y amor al saber.

A Ricardo Ciria, por ser un apoyo invaluable y una gran fuente de consejo.

A Alejandro Abdala, por ser un ejemplo de habilidad, dedicación y buen corazón.

A Esteban Peguero, por darme una imagen de un profesional a quien puedo aspirar a ser.

A más amistades, sin ningún orden particular: Teresa, Walter, Jimena, Juan, María Luisa, Alejandra, José Luis, Paola, Flavia.

Y a Aurora, por ser una persona increíble y sorprenderme al final de este capítulo con el más bello regalo que jamás pude haber recibido.

Resumen. Las células vivas obtienen las moléculas necesarias para mantenerse por medio de transformaciones mediadas por reacciones químicas, que forman parte del sistema conocido como metabolismo. Éste se organiza como una red metabólica, que se conforma de distintas rutas en las cuales los compuestos son transformados por pasos sucesivos. Los avances recientes de la genómica y la bioinformática se han usado para proponer métodos automáticos para analizar las redes metabólicas y encontrar patrones del funcionamiento del metabolismo. De estos métodos, los basados en la teoría de grafos organizan las redes metabólicas con base en las interacciones entre compuestos, obteniendo categorías funcionales que no conservan las rutas metabólicas. Otros métodos, basados en el intercambio de grupos químicos y el flujo de átomos, rastrean la transformación de los compuestos en detalle, pero requieren información del entorno en el cual sucede la reacción para asociar reactivos y productos. Las asociaciones procedentes de ambos tipos de métodos pueden ser organizadas con métodos matemáticos de agrupamiento para obtener grupos de reacción que describen transformaciones químicas similares y que pueden relacionarse con clases enzimáticas o rutas metabólicas.

Con el objetivo de hacer un método nuevo que requiera un mínimo de información contextual de las reacciones para clasificarlas de manera natural, este trabajo presenta un enfoque basado en reglas que agrupa las reacciones del metabolismo a través de las transformaciones entre sustratos y productos. Cada reacción fue organizada como una estructura de árbol (TS) de pares de compuestos y compuestos aislados. Los pares en las TSs se compararon con los reportados en la colección RPAIR de KEGG, con la cual se obtuvo una concordancia del 81 %. Las TSs agruparon naturalmente las reacciones en 71 patrones generales de uso de compuestos. La catálisis dentro de estos grupos fue evaluada con las categorías de la Comisión de Enzimas, revelando un uso preferencial de clases de enzimas por grupo.

Estos resultados muestran que con reglas simples se pueden hallar patrones que reflejan las transformaciones de pares sustrato-producto y el tipo de catálisis involucrado. Este enfoque puede usarse para generar inferencias sobre la catálisis de las reacciones, puesto que en el conjunto probado las reacciones carentes de un número EC se agruparon en TSs junto con aquellas que sí lo tienen. Este uso preferencial de las enzimas dentro de los grupos de TSs abre la posibilidad de emplear este enfoque como una clasificación independiente y de estudiar las relaciones evolutivas de las enzimas en cada grupo de TSs.

Summary. Living cells obtain the molecules they require by transforming them through chemical reactions, which are part of a system known as metabolism. This set is organized as a metabolic network made up of pathways, in which compounds are transformed in successive steps. Recent advances in genomics and bioinformatics have been used to propose automated methods to analyze metabolic networks and identify their functional patterns. Those methods based on graph theory organize the network through interactions among compounds, obtaining functional categories which fail to preserve pathways. Other methods, based on chemical group exchanges and atom flow, trace compound transformations in detail, but require information of the environment in which the reaction takes place to associate substrates to products. The associations obtained from both method types can be organized with mathematical grouping methods to generate reaction clusters that describe similar chemical transformations and can be related to enzyme classes or metabolic pathways.

With the goal of creating a novel method which requires a minimum of context information about reactions to classify them naturally, this work presents a rule-based approach which groups metabolic reactions through their substrate-product transformations. Each reaction was organized as a tree structure (TS) of compound pairs and loner compounds. The pairs in the TSs were compared with those with those reported in KEGG's RPAIR collection, obtaining a concordance of 81%. The TSs naturally clustered the reactions tested into 71 general compound usage patterns. The catalysis within these clusters was evaluated using the Enzyme Commission's categories, revealing a preferential usage of enzyme classes in each cluster.

These results prove that simple rules can be used to find patterns which reflect substrate-product transformations and the type of catalysis involved. This approach can be used to generate inferences on reaction catalysis, since in the tested reaction set reactions without an EC number were clustered into TSs with reactions with EC numbers. This preferential usage of enzymes within TS groups opens the possibility of using this approach as a standalone reaction classifier and of studying evolutionary relationships among the enzymes in each TS cluster.

Índice general

I Estado del arte y motivación

1. Introducción	1
1.1. Generalidades del metabolismo	1
1.2. Clasificaciones y bases de datos	3
1.2.1. Bases de datos del metabolismo	5
2. Antecedentes	9
2.1. Modelos computacionales del metabolismo	9
2.2. Modelos basados en la teoría de grafos	9
2.3. Mapeo de compuestos	13
3. Motivación y plan de trabajo	18
3.1. Justificación	18
3.2. Hipótesis	19
3.3. Objetivos	19
3.3.1. Objetivo General.	19
3.3.2. Objetivos Específicos.	19

II Materiales y Métodos de organización y agrupamiento **21**

4. Reglas e implementación	22
4.1. Parámetros de selección	22
4.2. Reglas elegidas	24
4.2.1. Regla de balance	24
4.2.2. Regla de conteo	25
4.3. Implementación de las reglas	27
4.3.1. Asociación entre grupos de compuestos	27
4.3.2. Implementación algorítmica de las reglas	30
4.4. Aplicación recursiva	31

III	Resultados y Discusión	35
5.	Generación de las clases	36
5.1.	Agrupamiento de las TSs	36
5.2.	Características generales de las TSs	38
5.3.	Descripción de los datos utilizados para la generación de las TSs	38
5.4.	Consideraciones sobre los agrupamientos	39
6.	Exploración de las TSs	41
6.1.	Exploración estadística del método	41
6.1.1.	Análisis estadístico de pares	41
6.1.2.	Análisis de los CTSs mediante clases de pares sustrato- producto	45
7.	Exploración de los CTSs	49
7.1.	Los CTSs tienden a agrupar categorías enzimáticas	49
8.	Conclusiones y Perspectivas	57
8.1.	Conclusiones	57
8.2.	Perspectivas	57
A.	Datos suplementarios	A-1
A.1.	Clusters de estructuras de árbol (CTSs).	A-1
A.2.	Relación de CTSs con clases de ECs a tercer dígito.	A-12
	Glosario y Siglas	A-14
	Bibliografía	B-1

Índice de figuras

2.1. Tipos de red relevantes para Barabási.	10
2.2. La estructura de la red metabólica cambia con la representación. . .	12
2.3. Un recorte de <i>hubs</i> sobre el metabolismo puede sobre ajustar la red a la distribución libre de escala.	13
2.4. Representación de la base técnica del mapeo de compuestos.	14
4.1. Esquema general de la regla de balance.	24
4.2. Ejemplo de fallo para la regla de balance.	26
4.3. Idea general de la regla de conteo.	26
4.4. Algoritmo implementado para las reglas de balance y conteo.	28
4.5. Resumen gráfico del enfoque.	32
4.6. Concepto de recursión.	33
5.1. Similitud entre TSs de compuestos distintos.	37
5.2. Los CTSs representan configuraciones únicas.	40
6.1. La precisión del presente enfoque varía entre CTSs.	43
6.2. La precisión es mayor para las clases <i>main</i> y <i>cofac</i> de Kotera et. al.	46
7.1. Los CTSs concentran diferentes grupos enzimáticos generales.	50
7.2. Los CTSs están enriquecidos selectivamente en diferentes grupos de ECs.	53

Índice de tablas

1.1. Metabolitos <i>pool</i>	2
1.2. Ejemplos de números de la Comisión de Enzimas (ECs).	4
1.3. Bases de datos de metabolismo.	6
1.4. Colecciones de datos de información química de KEGG.	7
5.1. Representaciones en diferentes formatos para TSs y CTSs.	38
6.1. Datos de distribución Beta para el conjunto de TSs.	44
7.1. Sólo los ECs de las oxidorreductasas son mutuamente comparables.	55

Parte I

Estado del arte y motivación

Capítulo 1

Introducción

1.1. Generalidades del metabolismo

El metabolismo es el conjunto de transformaciones químicas de los compuestos en una célula. Su importancia radica en que provee a los organismos vivos de la energía que éstos necesitan y de los bloques de construcción de las macromoléculas que los conforman. Estas transformaciones continuas de los compuestos ayudan al organismo a mantener su homeostasis [5].

El metabolismo se categoriza según el tipo de transformación de los sustratos en dos tipos: el anabolismo, en el cual los compuestos simples son usados como componentes de las macromoléculas, y el catabolismo, en el cual las macromoléculas son degradadas para obtener energía y compuestos simples.

En el metabolismo, las reacciones reciben sus sustratos de otras reacciones y ceden sus productos a otras reacciones, lo cual ha llevado a representarlo principalmente como sucesiones de reacciones denominadas rutas metabólicas. Estas rutas, como el ciclo de los ácidos tricarbóxicos (TCA), llevan a cabo secuencialmente la síntesis y degradación de algunos de los compuestos que la célula requiere [19].

Un aspecto notable del metabolismo es su redundancia; algunos compuestos son utilizados con frecuencia en el catabolismo y el anabolismo. Estos compuestos son llamados metabolitos *pool* [31], los cuales se enlistan en la tabla 1.1. Algunos metabolitos *pool* son coenzimas, que asisten a las enzimas para favorecer las transformaciones en el metabolismo, como Adenosina-5'-trifosfato (ATP), Adenosina-5'-difosfato (ADP), Nicotina-adenina dinucleótido (NAD), Nicotina-adenina dinucleótido reducido (NADH), NAD-fosfato (NADP), NAD-fosfato reducido (NADPH), o Coenzima A (CoA). Otros metabolitos *pool* son Amonio (NH_4) y Ortofosfato (PO_4).

Tabla 1.1: **Metabolitos *pool***. Descripción de los metabolitos de mayor frecuencia en la red metabólica descrita en la base de datos KEGG. La tabla, de izquierda a derecha, describe la posición de mayor a menor del compuesto (columna 2), ordenado según el grado proporcionado en la columna 3 y su función en el metabolismo (columna 4). Tomado de [31].

Lugar	Nombre	Grado	Función
1	Agua (H ₂ O)	2213	Hidrólisis, hidratación
2	Protón (H ⁺)	1269	Bombas de protones, oxidorreducciones
3	Oxígeno (O ₂)	860	Aceptor de electrones
4	NADP	724	Coenzima; Aceptor de electrones
5	NADPH	721	Coenzima; Donador de electrones en el anabolismo
6	NAD	663	Coenzima; Aceptor de electrones en el catabolismo
7	NADH	655	Coenzima; Donador de electrones
8	ATP	466	Coenzima; Donador de energía
9	CO ₂	427	Producto de la oxidación, precursor de la fotosíntesis
10	PO ₄	393	Producto de hidrólisis de ATP, ADP y AMP
11	CoA	369	Coenzima; Donador de grupos acilo
12	ADP	333	Producto, hidrólisis de ATP Sustrato, síntesis de ATP
13	NH ₄	296	Fuente de nitrógeno Producto, catabolismo de aminoácidos/nucleótidos
14	Pirofosfato	286	Producto, hidrólisis de ATP
15	S-Adenosilmetionina (SAM)	245	Coenzima; Donador de grupos metilo
16	S-Adenosilhomocisteína	236	Subproducto de metilación por SAM
17	UDP	222	Coenzima; transferencia de hexosas
18	H ₂ O ₂	163	Oxidorreducciones

Lugar	Nombre	Grado	Función
19	2-oxoglutarato	158	Parte del TCA; Transferencia de NH ₄
20	AMP	158	Producto, hidrólisis de ATP/ADP; Sustrato, síntesis de ATP/ADP
21	Piruvato	151	Final de glicólisis y algunas rutas de aminoácidos (Ala, Cys, Ser)
22	AcCoA	136	Coenzima; donador de grupo acetilo
23	Ácido glutámico (Glu)	129	Transferencia de NH ₄ Intermediario/precursor para aminoácidos
24	Oxalacetato	43	TCA, gluconeogénesis Precursor de asparagina

Muchas reacciones del metabolismo son asistidas por coenzimas, una clase de moléculas no proteicas que donan grupos químicos. Las coenzimas son notables por ser modificadas al mínimo durante la reacción, por lo cual se les ha considerado, para algunos propósitos, como análogas a los cofactores [26], que son normalmente átomos que donan o reciben electrones durante la reacción (Fe, Mn, Mg...). Se sabe que las coenzimas son necesarias para iniciar el flujo del metabolismo [27], y que su falta causa muerte celular [11].

Para los propósitos de estudios teóricos, esta redundancia ha hecho que las coenzimas y los metabolitos *pool* reciban tratamientos distintos cuando se trata de predecir nuevas rutas metabólicas, caracterizar flujos del metabolismo u obtener propiedades que ayuden a agrupar o clasificar rutas y reacciones del metabolismo. Estos tratamientos se discuten en secciones posteriores.

1.2. Clasificaciones y bases de datos

Muchos trabajos han tenido como objetivo generar clasificaciones o agrupamientos para determinar propiedades no evidentes del metabolismo, a fin de producir conocimiento para estudios de la evolución de éste y sus componentes [43, 30]. Otra forma, más usual, de generar una clasificación es agrupar los compuestos en el metabolismo por sus propiedades para encontrar rasgos

comunes entre los compuestos que participan en la catálisis que sirvan para encontrar rutas, como en el caso de clasificar los compuestos como coenzimas [5] o metabolitos *pool* [31]. Esta idea de clasificar los componentes del metabolismo se ha aplicado también para generar grupos útiles de reacciones y enzimas; uno de los más socorridos ha sido el esquema de números de la Comisión de Enzimas, o Números EC (*Enzyme Commission Number (EC)*) [44], que se sigue usando como referencia para buscar funciones enzimáticas. Los ECs están estructurados como un esquema jerárquico que describe desde categorías enzimáticas generales hasta enzimas individuales. Los ECs dividen las enzimas en seis clases principales, que son subdivididas en categorías de cuatro números, separados por puntos. La tabla 1.2 presenta ejemplos de cada una de las clases principales y de algunas subclases.

Tabla 1.2: **Ejemplos de números de la Comisión de Enzimas (ECs)**. Las seis grandes jerarquías de los ECs, con sus respectivos nombres y un ejemplo de cada una.

Jerarquía Primer dígito	Clase de enzima	Ejemplo	Nombre de la Enzima
1	Oxidorreductasas	1.1.5.9	Glucosa deshidrogenasa
2	Transferasas	2.3.1.1	Glutamato N-acetiltransferasa
3	Hidrolasas	3.1.1.2	Fenilacetato acetilhidrolasa
4	Liasas	4.1.2.13	Fructosa- difosfato aldolasa
5	Isomerasas	5.4.2.2	Fosfoglucomutasa
6	Ligasas	6.3.1.1	Aspartato-amonio ligasa

Cada dígito representa el lugar de una enzima en la jerarquía. El nivel más general es el primer dígito, a la izquierda, que representa las clases generales que se ejemplifican en la tabla 1.2.

1.2.1. Bases de datos del metabolismo

El metabolismo es un sistema vasto, de miles de reacciones. Éstas han incrementado con la información proveniente de los métodos de la genómica y de la compilación de datos de grandes cantidades de experimentos no hechos a gran escala. Para asistir el trabajo relacionado, se ha hecho indispensable generar bases de datos para manejar esta información.

Las bases de datos, más allá de almacenar la información, organizan y relacionan los datos para facilitar su acceso e integración. Existen varias de ellas, con estructuras e información variables que en ocasiones son complementarias. La tabla 1.2.1 ilustra algunas de las más utilizadas, cuya información puede aprovecharse para construir modelos del metabolismo.

Para este trabajo, se escogió la Enciclopedia de Genes y Genomas de Kioto (*Kyoto Encyclopedia of Genes and Genomes* (KEGG)), Japón [22]. Este compendio concentra una variedad amplia de datos del metabolismo y secuencias de más de 4000 genomas completos, así como varias herramientas para trabajar estos datos y ponerlos en contexto, con un enfoque inclinado a la medicina y la salud.

El núcleo de la organización de KEGG es la base de datos de ortología (*KEGG Orthology, KO*) [21], que liga las secuencias disponibles a funciones generales respecto de GENES, una colección de datos sobre genes que originalmente se limitaba a los identificados en genomas completos. KO funciona acumulando ligas a estos genes; los identificadores de KO concentran la función que comparte un conjunto de los genes ortólogos en GENES, independientemente del organismo en el cual se encuentren [22]. Esto ha permitido usar los identificadores de KO para hacer varios sistemas de anotación con un éxito significativo [6, 33, 34].

La categoría de datos de KEGG más relevante para este trabajo es la categoría de información química. Ésta empezó como una colección llamada LIGAND, que eventualmente se expandió para abarcar conocimiento sobre el universo de sustancias químicas y reacciones relevantes para la vida, particularmente sobre el metabolismo celular. La tabla 1.4 expone de forma resumida el contenido de estos conjuntos: COMPOUND, para compuestos metabolizables y relacionados; ENZYME, para enzimas, anotadas según los ECs; REACTION, que asocia enzimas con los compuestos que usan; y RCLASS, que asocia a los compuestos por la transformación por la que atraviesan en una reacción. Cabe mencionar aquí que COMPOUND y REACTION presentan información sumamente detallada en términos de la composición de los compuestos y la forma en que éstos pueden participar en una reacción [22]. La organización de KEGG, como muestra la tabla, da identificadores a cada componente que permiten manejarlo de forma única y directa, lo cual

Tabla 1.3: **Bases de datos de metabolismo**, comparadas por sus objetivos y características. Cuadro comparativo de las bases de datos mayormente utilizadas en estudios bioinformáticos.

		ChEBI	MetaCyc	KEGG
	Contenido	compuestos	general	general
Datos	Estructura	ontología	liga general	liga general
	Propiedades	clasificación jerárquica por tipo; roles ligan a funciones	curación manual extensiva; tipos de dato definidos y separados	información externa en el mismo archivo; centralizados por KEGG Orthology
Enfoque	Objetivo	desarrollar relaciones entre componentes	dar información para modelar datos con validación experimental	asignación de significados funcionales
	Medios	componente principal de otras fuentes de datos (GO, Uniprot)	Software integrado (Pathway Tools); Incrementar cobertura de datos	Contextualizar por mapeo a rutas; corrección por agregado de identificadores
Usos		fuentes de identificadores para bases de datos externas; minería de datos	modelado de organismo completo; integración de datos de experimentos	análisis de high-throughput; metagenómica; búsqueda de información médica para el público

facilita el trabajo bioinformático.

Cuando KEGG fue iniciada, los ECs en ENZYME eran usados como identificadores primarios para construir rutas metabólicas sobre el genoma

Conjunto de datos	Tipo	Identificador único
COMPOUND	Compuestos	C00001
REACTION	Reacciones	R00002
GLYCAN	Poliazúcares	G00001
ENZYME	Enzimas	EC (ej.: 1.1.1.1)
RCLASS	Pares de compuestos	Par (ej.: C00001_C00008)

Tabla 1.4: **Colecciones de datos de información química de KEGG.** Se muestra la organización de los datos de LIGAND y de algunas de sus tablas, así como ejemplos de los identificadores correspondientes a cada tabla.

completo, trazando mapas de las rutas que usaban a los ECs como nodos. En 1999, se introdujeron identificadores de ortología como sustituto de los ECs, para cubrir elementos dentro y fuera del metabolismo. En 2002, estos identificadores de ortología fueron estructurados como KO. Actualmente, los ECs son tratados como atributos de KO relacionados con las reacciones asociadas a rutas metabólicas. Este trabajo aprovecha esta relación para asociar reacciones con enzimas.

Una particularidad de REACTION es que no asigna un significado específico a los lados de cada reacción, pudiendo ser cualquiera de ellos el de los sustratos o el de los productos. En consecuencia, por ejemplo, el par $\text{ATP} \leftrightarrow \text{ADP}$ aparece en ese orden en la reacción R00760 y como $\text{ADP} \leftrightarrow \text{ATP}$ en la reacción R00140.

RCLASS surgió de un conjunto de datos previo, denominado RPAIR, el cual pretendía dar información de transformaciones de pares de compuestos como RCLASS, pero también en el contexto de la participación de cada par de compuestos en la reacción. RPAIR procede de una clasificación de pares de reactantes propuesta por Kotera et al., orientada a revelar el rol de cada par en una reacción: *main* (cambios en sustratos/productos principales), *cofac* (cambios en coenzimas para oxidorreductasas), *trans* (grupos transferidos en transferasas), *ligase* (uso de nucleósidos-trifosfato en ligasas), y *leave* (separación o adición de compuestos inorgánicos en enzimas como liasas e hidrolasas) [26]. RPAIR fue eliminado para conservar RCLASS, del cual se argumenta que contiene los pares mejor curados. Sin embargo, RCLASS carece de la clasificación original de RPAIR.

Al igual que REACTION, RPAIR/RCLASS es indiferente a la direccionalidad de la reacción, pero RPAIR/RCLASS lo hace por motivos de representación; cada uno de los pares es representado siempre empezando por el identificador de COMPOUND más chico. Por ejemplo, un par de reducción de Flavina mononucleótido (FMN), $\text{FMN} \leftrightarrow \text{FMN reducido}$ (C00061_C01847)

siempre se representa en RPAIR en ese orden, sin importar si FMN está del lado derecho, como en la reacción R05218, en vez del izquierdo, como en la reacción R07664. Lo mismo puede decirse del par ATP \leftrightarrow ADP (C00002_C00008).

La organización antes descrita, la disponibilidad de descripciones detalladas de las reacciones, y la presencia de pares sustrato–producto en el conjunto RCLASS hicieron a KEGG el compendio adecuado para el desarrollo de este trabajo.

Capítulo 2

Antecedentes

2.1. Modelos computacionales del metabolismo

Para poder desarrollar modelos del metabolismo, es necesario un trabajo de curación sobre los datos disponibles para determinar cuales pueden usarse para un modelo completo y de qué formas. La información contenida en las bases de datos ayuda a completar y detallar los modelos que pueden hacerse, pero es problemático manejar cantidades de datos tan grandes de forma exclusiva y manual. A fin de aprovechar esta información, se han desarrollado varios métodos de cómputo para asistir la labor de curación y el desarrollo de modelos.

2.2. Modelos basados en la teoría de grafos

Uno de los enfoques de mayor impacto para aplicar métodos de cómputo al estudio de sistemas biológicos fue la teoría de grafos, puesto que permitió cubrir conjuntos de datos muy grandes con recursos de cómputo modestos representándolos como redes de interacciones (ver grafo en Glosario). Los primeros enfoques empezaron por cubrir la totalidad de la célula, a fin de determinar propiedades generales que permitieran modelar el comportamiento global del organismo.

Una de las primeras aplicaciones de este tipo fue la impulsada por Albert Lázlo Barabási [3]: tomar los compuestos como nodos y examinar sus interacciones para observar propiedades generales. Barabási ya había tenido un éxito notable aplicando este enfoque a redes de regulación de la transcripción centrándose en la relación entre genes reguladores y regulados. Posteriormente, Barabási trasladó esta idea al metabolismo; al tomar a los compuestos como el componente principal del sistema (y por ende los nodos), embebió las inter-

acciones entre compuestos (la acción de las enzimas) en las aristas, colocando una entre dos compuestos si había alguna enzima que usara ese compuesto, ya fuera como reactivo o como producto.

Al trabajar con redes de regulación, Barabási había observado un acomodo jerárquico–modular, que se ilustra en la figura 2.1: una red jerárquica con pocos nodos en las posiciones superiores, pero que conectaban muchos otros nodos. Estos nodos superiores fueron nombrados *hubs*, distinguibles por el hecho de interconectar gran cantidad de elementos del sistema. Los nodos de grados menores presentaban una propiedad interesante: tendían a asociarse en grupos que conectaban más a sus nodos integrantes con otros nodos del grupo que con los de grupos externos. Estos grupos fueron denominados *módulos*: los genes del módulo se podían identificar claramente como miembros de éste. Barabási propuso que esta organización puede reconocerse por propiedades generales de la red dadas por una distribución de los grados. La distribución de grados ($P(k)$) representa el número de nodos que tienen un cierto grado en la red, y la distribución de clustering ($C(k)$) hace esto mismo para los coeficientes de clustering (ver Glosario). Para Barabási, en una red jerárquico–modular ambas distribuciones se ajustan a una ley de potencia – una exponencial de exponente negativo [20]; hay pocos nodos conectados con muchos otros nodos, y muchos nodos conectados con pocos.

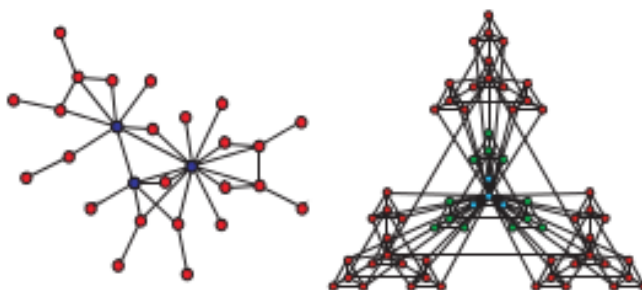


Figura 2.1: **Tipos de red relevantes para Barabási.** Izquierda: red libre de escala. Derecha: red jerárquico–modular. Tomado de [40].

En el primer intento de observar las propiedades de la red, Barabási encontró una red con un módulo gigante o megamódulo, en el cual se agrupaban muchas reacciones que no estaban funcionalmente relacionadas. Para examinar las posibles propiedades de la red, Barabási eliminó compuestos que pueden considerarse *hubs* — precisamente, los metabolitos *pool* — y observó el resultado de la descomposición. Ésta reveló varios módulos dentro de la red, pero su organización no era estrictamente jerárquico–modular; los

módulos eran menos obvios de lo que hubiera cabido esperar tomando el patrón observado para redes de regulación. Este modelo fue llamado *libre de escala* (*scale-free*), también ilustrado en la figura 2.1, con una $P(k)$ lineal y una $C(k)$ de ley de potencia. El sólo hecho de organizarse como una red libre de escala con módulos, de acuerdo con Barabási, refleja propiedades ya observadas en otras redes libres de escala, como tolerancia a la inclusión o eliminación aleatoria de nodos [20].

De igual forma, Barabási había subrayado la importancia de los módulos a un nivel más directamente biológico, puesto que los módulos que encontraba en redes biológicas traslapaban fuertemente con funciones biológicas generales, como sectores de manejo de nucleótidos [40]. Esta observación se confirmaría después usando redes proteína-proteína de varias fuentes de datos [32]. Barabási había logrado identificar estas secciones de redes biológicas usando un proceso estrictamente matemático, lo que atrajo a este tipo de métodos inmediata atención.

El éxito del método de Barabási llevó a ampliar los esfuerzos de investigación en el tema de redes biológicas, incluyendo el caso de trabajos enfocados a investigar las propiedades del método para solidificar su respaldo y hallar posibles mejoras. Este tipo de trabajos encontraron diversas características de la hipótesis de la red libre de escala que no se ajustaban a la estructura de la red metabólica. La discusión subsecuente produjo que estudios futuros se enfocaran en otras propiedades del metabolismo.

Una de las críticas más importantes de la red libre de escala fue desarrollada en el trabajo de Masanori Arita. Uno de sus artículos toma el número EC como base y delinea el intercambio de grupos para ubicar las correspondencias de átomos entre sustratos y productos con ayuda de un método semiautomático [1]. Su conclusión al analizar las propiedades del conjunto de datos generado es que las redes metabólicas no cumplen con muchos supuestos clave que los métodos de grafos suponían *a priori*, como la libertad de escala. Arita también señaló puntos relevantes sobre la representación de los datos, como el hecho de que distintas representaciones de la red proveen distinta información y arrojan distintos resultados con los mismos métodos topológicos, puesto que cada representación refleja distinta información [2]. Por ejemplo, un grafo de metabolismo puede tomar como elementos centrales únicamente a los compuestos, o puede tomar tanto compuestos como reacciones, lo cual lo hace lo que se conoce como un grafo *bipartita*; los nodos representan dos tipos de objeto distintos que se manejan como similares en la red para ver sus relaciones. No obstante, la representación del grafo bipartita tiene propiedades estadísticas distintas a las de un grafo de compuestos, puesto que el acomodo de los nodos cambia. La figura 2.2 presenta un ejemplo de este caso con base en la glicólisis: un grafo de compuestos tiene una

estructura diferente a la de un grafo bipartita, pese al hecho de ambos representan el mismo sistema. Por ejemplo, la síntesis de gliceraldehído 3-fosfato desde la fructosa-1,6-bifosfato aparece en el grafo de compuestos como un triángulo, mientras que el grafo bipartita la presencia de las enzimas da una topología distinta. Pese a que ambos representan el mismo sistema, los acomodos de los nodos son diferentes, y la acumulación de las diferencias cambia sustancialmente las propiedades matemáticas de ambos grafos.

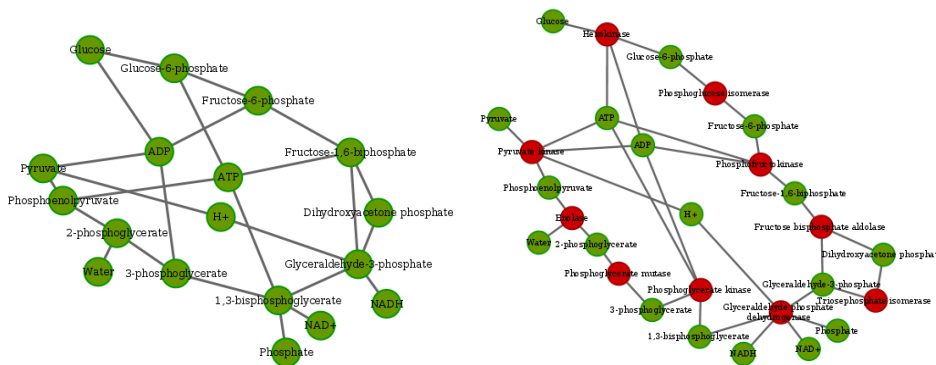


Figura 2.2: **La estructura de la red metabólica cambia con la representación.** Representaciones de la glicólisis. Izquierda: grafo de compuestos. Derecha: grafo bipartita de compuestos y reacciones.

Las críticas a los métodos topológicos fueron creciendo en la literatura, hasta llegar a los supuestos mismos del enfoque subyacente. Un ejemplo interesante vino de Gypsi Lima Méndez y Jacques van Helden [31]. Al examinar con cuidado los datos y el método de Barabási, se encontraron diversas fallas. La más importante fue el modo en que se llegó a la distribución final. La red original no sigue la distribución libre de escala que Barabási proponía; la red había sido recortada para coincidir con la distribución libre de escala, como se ilustra en la figura 2.3. Trabajo previo ya había indicado que al remover los *hubs*, como agua o ATP, la red se fragmenta revelando la distribución libre de escala esperada [13]. La distribución tomada sin retirar nodos tiene características que debilitan el ajuste a una ley de potencia, como una desviación entre los nodos de bajo grado; el ajuste mejora al eliminar los 30 nodos de mayor grado [31]. Esta desviación entre los nodos de bajo grado incluso es apreciable en un artículo de Barabási (Figura 2.3F), pero en éste se considera que el ajuste es adecuado [40]. De igual forma, la falta de ajuste a la distribución lleva a la ausencia de las propiedades propuestas como la tolerancia a fallos, puesto que muchas se deben a aristas poco significativas entre metabolitos de alto grado o implicarían la eliminación de cientos de enzimas, como la idea de eliminar el agua como nodo [31].

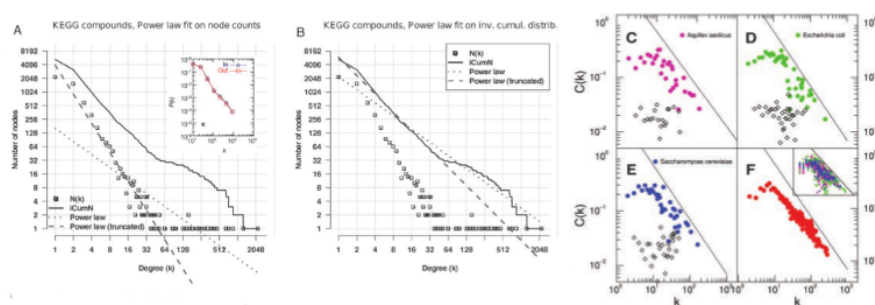


Figura 2.3: Un recorte de *hubs* sobre el metabolismo puede sobre ajustar la red a la distribución libre de escala. A,B: Ajuste de la distribución de ley de potencia a la distribución de la red de compuestos de KEGG proporcionado por Lima en [31]. Abcisa: grado; Ordenada, número de nodos con el grado dado por la abcisa. A, cuentas de nodos; B, distribución acumulada inversa. C–F: ajuste a la ley de potencia de distribuciones de organismos proporcionada por Barabási como evidencia. Abcisa, grado; Ordenada, clustering promedio de los grados dados por la abcisa. C, *Aquifex aeolicus*; D, *Escherichia coli*; E, *Saccharomyces cerevisiae*; F, el promedio de los 43 organismos analizados en [40], con todos los datos en el recuadro. A y B tomados de [31]; C a F tomados de [40].

Debido a estos resultados, los métodos de redes basados en el modelo libre de escala perdieron relevancia para estudios del metabolismo, para dar lugar a métodos alternos para determinar propiedades de las rutas metabólicas. Hoy se sabe, por ejemplo, que la topología de las redes es por sí sola insuficiente para revelar la dinámica que provoca cambios de estado en la red, como la activación o represión de genes [12]. El trabajo subsiguiente se concentraría en las propiedades de las transiciones químicas para ayudar a la predicción de rutas.

2.3. Mapeo de compuestos

Con base en las críticas a los métodos topológicos para el metabolismo, las nuevas familias de métodos se centraron en las características de los compuestos, útiles para ubicar una ruta metabólica al examinar reacciones y rutas. Estos métodos se basan en la transformación de los compuestos en cada reacción, como en la propuesta de Arita [1]. Muchos grupos se dieron a la tarea de automatizar este tipo de comparaciones para asistir el esfuerzo de curación manual. Este enfoque dio origen a los métodos de mapeo de compuestos, que son actualmente preferidos sobre los métodos de teoría de

grafos para trazar reacciones y rutas en el metabolismo.

La idea central del mapeo de compuestos se basa en el análisis de la estructura de los compuestos. Al comparar el acomodo de los átomos en la estructura de los compuestos de forma directa, es posible detectar los grupos químicos que los distinguen, y que por ello reflejan la transformación que se lleva a cabo entre pares sustrato-producto. Esto permite comparar compuestos en lados opuestos de la ecuación y detectar el grupo que se pierde o se adquiere. Este enfoque es atractivo porque es intuitivo, señala los grupos químicos que se detectan y se le puede dar también poder predictivo [42]. La idea general se ilustra en la figura 2.4.

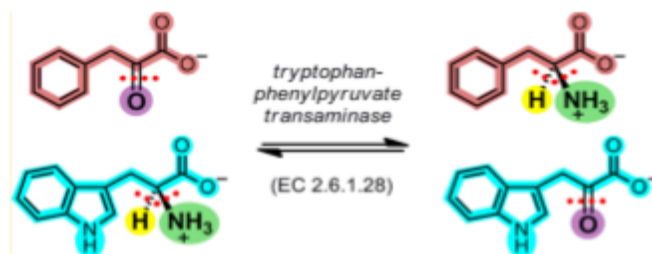


Figura 2.4: **Representación de la base técnica del mapeo de compuestos.**

La estructura de cada compuesto se abstrae como una representación de una imagen, que permite a una computadora compararlo con otros. El mapeo ubica las diferencias buscando qué partes de la estructura de un compuesto se parecen a las del otro (rojo con rojo, azul con azul), lo cual permite comparar los intercambios de grupo directamente. Tomado de [29].

Un ejemplo de implementación de un mapeador de compuestos es SIMCOMP, un programa usado por KEGG para proveer datos a RPAIR [42]. SIMCOMP se concentra en encontrar compuestos similares a un compuesto indicado para búsqueda. SIMCOMP usa primariamente dos parámetros para aproximar el resultado, basados en tiempo de cómputo y fragmentos completos de compuestos [14]. Cuando SIMCOMP fue liberado, era un enfoque novedoso que le dio ventaja sobre otros enfoques de ese momento, sobre todo los basados en cadenas de bits [14]. SIMCOMP puede aceptar cualquier compuesto como entrada, lo cual lo orienta naturalmente a la exploración de rutas [15].

Un enfoque alterno ha sido reforzar los métodos topológicos tradicionales con supuestos adicionales compartidos con los métodos de mapeo de compuestos. Se han sintetizado las observaciones de Arita y otros bajo el supuesto de que *una ruta de interés debe transferir al menos un átomo del inicio al final*. Usando esta observación, se desarrolló un método llamado ReTrace,

que busca simplificar el trazado de bifurcaciones en rutas de interés usando un método de aproximación. ReTrace encuentra estas rutas eliminando la necesidad de mostrar metabolitos laterales a la ruta de interés o compuestos que no contribuyen al flujo de carbono [37].

Los métodos disponibles de mapeo de compuestos se han conjugado con las metodologías de análisis de balance de flujos (*flux balance analysis* (FBA)). El FBA surgió con el objeto de describir la distribución del flujo metabólico de forma cuantitativa [45], y eventualmente ha permitido rastrear rutas a través de sistemas de gran tamaño. El FBA modela todas las reacciones del sistema como ecuaciones de sustratos y productos, incluyendo una *función objetivo*, que es la transformación que el modelo debe optimizar; el objetivo del FBA es usar las reacciones disponibles para producir el flujo máximo posible a través de la función objetivo [8]. El FBA es visto como un método simple y extensible, puesto que requiere sólo datos de estequiometría y de equilibrios de masa de los metabolitos; una vez se afina el modelo, los flujos metabólicos y las rutas pueden resolverse con matrices mediante programación lineal [28]. Uno de los supuestos clave del FBA es el llamado *supuesto del estado pseudo-estable*; la expectativa de que la dinámica estequiométrica de una ruta metabólica es constante en su interior, debido a que cada compuesto se produce y consume a la misma tasa. Por esta razón, se asume que la estructura estequiométrica del metabolismo completo basta para llevar al organismo a una homeostasis [23].

A medida que se han agregado parámetros fisicoquímicos y alternativas para optimizar el desempeño, los métodos de mapeo de compuestos y rastreo de átomos han seguido evolucionando desde su potencial para revelar rutas más coherentes [10] por la minimización de conexiones poco informativas [31]. Algunos son *optStoic*, que permite trazar eficazmente los flujos de carbono con la inclusión de parámetros termodinámicos [7]; y DESHARKY, que intenta predecir la viabilidad de una ruta dada en un organismo a partir de los parámetros de flujo existentes [41]. Otro ejemplo es el mapeador de compuestos desarrollado para la *suite* de herramientas Pathway Tools de la base de datos BioCyc. Este mapeador está basado en programación lineal de enteros con mezcla, que permite representar los valores en las matrices con decimales. Esto, de acuerdo con los autores, permite representar los resultados de forma más eficiente, lo cual hace a este método uno de los más rápidos hoy disponibles [29].

El mapeo de compuestos resulta una idea muy intuitiva para resolver compuestos y rutas. No obstante, tiene una limitación importante; las computadoras tienen dificultad para procesar imágenes. Se han logrado avances importantes en áreas de reconocimiento de imágenes, pero las computadoras aún requieren puntos de referencia previamente señalados para poder

reconocer un objeto.

Para resolver esto, el acercamiento que usan los mapeadores de compuestos es abstraer los enlaces entre los átomos del compuesto para representarlos como un grafo (figura 2.4). Para poder manejar el grafo en un dispositivo de cómputo, conviene representarlo como una matriz; en el caso de un mapeador de compuestos, la matriz representa los enlaces que se dan entre los átomos que conforman el compuesto [17]. De esta forma, los puntos de referencia básicos son las posiciones relativas de los átomos. La matriz puede complementarse haciendo lo que se llama *colorear el grafo*; señalar el elemento que corresponde a cada átomo y los tipos de enlaces que comparten. No obstante, usar el grafo implica una dificultad computacional: resolver variantes del *problema de isomorfismo de subgrafos* [17] (ver Glosario). Actualmente se sabe que, en general, este problema está dentro del tipo NP-completo (ver Glosario), lo cual implica dificultades de implementación. Una estrategia frecuente para evitar estos problemas es forzar un límite arbitrario de tiempo de ejecución y, si se alcanza el límite, curar a mano [17].

Se han propuesto intentos de mejora basados en la modificación de la lógica del algoritmo, para que resuelva, por ejemplo, el problema de isomorfismo de grafos, un problema relacionado para el cual hay casos particulares que son rápidos [17]. Sin embargo, el problema reside en que muchos compuestos de interés no pueden trabajarse satisfactoriamente con esta modificación, y el proceso sigue siendo muy exigente computacionalmente pese a la mejora. Igualmente, hay trabajo que apunta a que resolver rutas metabólicas en general es un problema NP-duro [36], lo cual hace la solución general computacionalmente impráctica, y que es necesario incluir datos que aclaren la dinámica del sistema [12]. Por estos motivos, los métodos combinados de FBA-mapeo tomaron relevancia sobre los de mapeo de compuestos exclusivo.

La mayoría de los métodos disponibles mitigan las limitantes incluyendo complementos para incrementar la velocidad y concentrar el espacio de búsqueda del algoritmo. El mapeador de Pathway Tools excluye coenzimas conocidas y usa heurísticos para compuestos con anillos [29]. DESHARKY ignora compuestos que no son “específicos al organismo hospedero” [41]. *optStoic* incluye varios pasos de optimización, a fin de constreñir posibles reacciones en términos de equilibrio de masas y cargas, factibilidad termodinámica, e incluso costos monetarios de compuestos, entre otros factores [7]. Estas operaciones incrementan la rapidez e incluso la precisión con la cual los métodos disponibles pueden identificar transiciones de interés, pero también complican los métodos subyacentes.

La mayor parte de las descripciones de limitantes de los procesos de FBA-mapeo se han concentrado en el manejo de fuentes alternas de datos, como regulatorias [45] y ecológicas [46]. No obstante, se han empezado a encontrar

rutas metabólicas factibles que no siguen el supuesto del estado pseudoestable, y que por lo tanto no contribuyen a conducir a una homeostasis de forma automática; algunos ejemplos son la ruta de las pentosas fosfato, el TCA y la síntesis de arginina [38]. Esto ha llevado a esfuerzos para nuevas mejoras; uno de ellos, AGPathFinder, incluye parámetros termodinámicos, esquemas de pesado, y formas de evadir los metabolitos *hub* [19]. Cabe señalar que lo hace rastreando cada interacción simplificando los compuestos a lo que llaman “grupos atómicos” basados en la vecindad por enlaces covalentes, a fin de trazar la transición completa como un grafo dirigido. El algoritmo es poderoso, capaz de rastrear posiciones de átomos en los compuestos de una ruta y demostrablemente eficaz frente a otros métodos, pero también usa métodos de simplificación no triviales y requiere recursos de cómputo considerables.

No obstante, un factor crucial que se menciona poco es el argumento de evitar las transiciones “poco significativas” [19, 31]. La idea es que los primeros métodos de grafos no tomaban en cuenta la acción de enzimas específicas, puesto que todas las interacciones entre un par de compuestos quedaban embebidas dentro de la misma arista. Una consecuencia de esto es que no se discriminaba entre transiciones; es posible hacer, por ejemplo, ATP de etanol en dos pasos, por el flujo de un grupo hidroxilo por agua. Este tipo de transiciones químicas son posibles, pero no son suficientemente informativas para trazar una ruta metabólica [31].

Una forma de manejar la importancia relativa de las transiciones es eliminar las coenzimas, con el argumento de que la transición entre las moléculas del par coenzima ya es conocida [29]. El uso de las coenzimas, no obstante, no es únicamente donar o recibir grupos químicos; una coenzima puede transformarse de otras formas en otras reacciones. Ejemplos son las reacciones de síntesis de las coenzimas, en las cuales el producto es la coenzima misma, y el uso de ATP en la elongación de ácidos nucleicos complejos. Esto se da también entre otros compuestos que no son considerados coenzimas. El par Glutamina (Gln) \leftrightarrow Glu, por ejemplo, está más asociado con la síntesis de proteínas, pero es uno de los donadores de nitrógeno, bajo la forma de NH_4 , más importantes del metabolismo [18]. En KEGG, el par Gln \leftrightarrow Glu aparece con frecuencia en un número importante de reacciones desempeñando este rol.

Capítulo 3

Motivación y plan de trabajo

3.1. Justificación

Uno de los principales intereses en el campo de la Bioinformática y la Genómica es analizar las redes metabólicas con el fin de extraer propiedades útiles para desarrollar modelos cuidadosos que aumenten nuestro conocimiento biológico. Algunos de los métodos que han tratado de resaltar estas propiedades se basan en la teoría de grafos, que muestra tendencias generales entre los compuestos y reacciones del metabolismo que puedan dar información desde la estructura del sistema. Sin embargo, la representación global del sistema es insuficiente para dar información que vaya más allá de propiedades estadísticas generales, que no son fáciles de traducir a propiedades de rutas metabólicas específicas. Una interpretación profunda requiere de un análisis que permita cubrir todos los datos y sus propiedades para reconstruir e interpretar el sistema, desde tomar las características de amplio espectro hasta examinar reacciones caso por caso.

Algunos autores han demostrado que esto puede hacerse incorporando conocimiento bioquímico al concepto de red: las vías relevantes pueden deducirse siguiendo los intercambios de grupos químicos entre compuestos o descomponiendo cada reacción en pares sustrato–producto. Estos métodos han logrado llegar a niveles de detalle que incluso llegan al átomo al predecir el flujo en una ruta metabólica. Sin embargo, las limitaciones de cómputo inherentes a estos métodos obligan a omitir *a priori* transiciones de compuestos de alta frecuencia como las que se dan por las coenzimas, lo cual impide usar estos compuestos para hallar rutas nuevas y elimina las propiedades que agregarían a un análisis del metabolismo completo.

A fin de desarrollar un modelo que pueda usarse para determinar propiedades de reacciones y rutas metabólicas, es necesario minimizar la cantidad

de información del contexto que requiere la elaboración de un modelo, a fin de las propiedades de los pares sustrato–producto puedan usarse para el modelado sin necesidad de complementarlas con procesos adicionales. Con esta visión, en este trabajo se presenta un modelo basado en reglas fundamentadas en las propiedades de las reacciones del metabolismo para crear un nuevo modelo de clasificación de reacciones que agrupa eventos catalíticos similares de manera natural. Este agrupamiento ofrece un esquema de análisis de las reacciones que puede usarse de forma independiente, puesto que proporciona conjuntos de pares sustrato–producto fiables con sólo un par de reglas simples. Además, estos pares sustrato–producto pueden usarse como un complemento para algoritmos de búsqueda de rutas metabólicas y como un parámetro para hallar reacciones candidato a curación manual.

3.2. Hipótesis

La similitud en las transformaciones químicas de los compuestos deriva en reglas de organización que agrupan las reacciones del metabolismo.

3.3. Objetivos

3.3.1. Objetivo General.

Establecer reglas para determinar la similitud de los compuestos en cada reacción en el metabolismo y usar estas reglas para clasificar las reacciones.

3.3.2. Objetivos Específicos.

1. Revisar la literatura para buscar principios de similitud de compuestos.
2. Usar los principios para crear reglas de agrupamiento de compuestos.
3. Establecer un proceso para usar las reglas en cada reacción.
4. Usar el proceso para generar estructuras de grafo.
5. Agrupar los grafos obtenidos en clases por su similitud topológica.
6. Determinar la validez del enfoque comparando con una base de datos curada a mano.

7. Comparar la clasificación de reacciones obtenida y el esquema de los números EC.

Parte II

Materiales y Métodos de organización y agrupamiento

Capítulo 4

Reglas e implementación

4.1. Parámetros de selección

Para construir el modelo, se buscaron en la literatura patrones de uso de compuestos en las reacciones reportadas en el metabolismo. Se notó que el patrón más común es el uso de coenzimas, que se usan preferentemente en tipos de reacción específicos; por ejemplo, las fosfotransferasas tienden a usar ATP y otros nucleósidos fosfato, las oxidorreducciones tienden a usar fuentes de poder reductor como NAD o Flavina–adenina dinucleótido (FAD), y las reacciones de síntesis de moléculas carbonadas tienden a usar CoA y moléculas de función similar.

Sin embargo, el rol de las coenzimas no es invariablemente el mismo; por ejemplo, el uso de una coenzima cambia en su reacción de síntesis. Por ende, una coenzima no puede eliminarse *a priori* sin perder la información de sus usos alternos. Esta idea revela que el rol de los compuestos en una reacción debe examinarse de forma *local*, tomando cada reacción por separado y distinguiendo los compuestos en esa reacción particular. Al mismo tiempo, la forma en que se examinan las reacciones debe ser idéntica para todas ellas, puesto que su base química es la misma.

Para desarrollar las reglas, se tomaron criterios relativamente simples para diferenciar entre compuestos, derivados de los siguientes supuestos de la estructura del metabolismo:

- *Las coenzimas sufren cambios mínimos en una reacción.* Los compuestos en una reacción tienden a sufrir cambios de apenas unos pocos átomos (comúnmente estos cambios son únicamente de grupos químicos relativamente pequeños y fijos), lo cual es coherente con el supuesto de la sección 2.3 – *una ruta debe transferir al menos un átomo del principio al final* – y con el hecho de que las rutas consisten en varias reacciones

sucesivas que modifican el compuesto de forma gradual, ambos expresados en la literatura como supuestos intuitivos [14, 16, 19]. Esto es muy notorio en compuestos que usualmente funcionan como coenzimas, como ATP y NAD, que se limitan a ceder grupos funcionales que son apenas fracciones minoritarias de su estructura total (ortofosfato en el caso de ATP y H^+ en el caso de NAD).

Los mapeadores de compuestos usan directamente las estructuras de cada compuesto para trazar los grupos cedidos. Los elevados tiempos y recursos de cómputo requeridos inclinan a no analizar las transiciones que involucran compuestos de estructuras complejas o alta frecuencia en el metabolismo, como muchas coenzimas. Para evitar esta dificultad, se eligió un modo de relacionar compuestos que no requiriera estructuras; seguir el patrón del mínimo cambio de las coenzimas usando la fórmula química como medida de similitud. A esta medida se le llamó de *balance* porque se basa en la diferencia de balance de masas entre los compuestos de una reacción; los compuestos que tienen la diferencia de masas más pequeña son definidos como los más similares en general. Un ejemplo de esto es el par NAD–NADH; la única diferencia entre ellos es un ión H^+ .

- *Las coenzimas están ampliamente difundidas en el metabolismo y son usadas en rutas poco relacionadas.* Tomando ATP y NAD como ejemplo, se nota que estos compuestos se usan para transferir energía entre reacciones, y por ende son clave para proveer esta energía a las rutas de anabolismo que la requieren. También se pueden encontrar en rutas de catabolismo, donde capturan la energía que las rutas anabólicas usan. Otros compuestos que siguen comportamientos similares son la CoA y el par Gln–Glu, que cede NH_4 ; los grupos son capturados del catabolismo y cedidos al anabolismo. Algunos pares de compuestos pueden tener diferencias de masa idénticas, lo cual hace imposible distinguirlos de esta forma; un ejemplo es la reacción R00756 (fosfofructocinasa: la fosforilación de fructosa-6P a fructosa 1,6-2P); la diferencia entre el par ATP–ADP y el azúcar es un grupo PO_4 , por lo que no es posible decir qué par es más cercano a una coenzima con los pesos moleculares. Para estos casos, se escogió una segunda medida, que toma el número de veces que un par de compuestos se repite en el metabolismo, siguiendo el patrón de la alta frecuencia global de las coenzimas. Esta propiedad fue usada por los métodos de redes para los compuestos individuales, pero no se ha usado antes para interacciones de compuestos, que reflejan más de cerca las transiciones químicas. A esta medida, por ser una cuenta simple, se le llamó de *conteo*.

4.2. Reglas elegidas

4.2.1. Regla de balance

La regla de balance busca encontrar las relaciones sustrato–producto, comparando grupos de combinaciones entre sustratos y productos. Para ello, usa como medida de comparación el intercambio de masas que ocurre entre los compuestos en una reacción balanceada, con coeficientes y subíndices conocidos, y con compuestos con pesos moleculares claramente definidos.

Un esquema general de la regla de balance se presenta en la figura 4.1. En el ejemplo toma una reacción de deshidratación catalizada por la enolasa, y se busca encontrar una diferencia entre los pesos moleculares de los compuestos que distinga a los compuestos principales para asociarlos como sustrato y producto de la reacción. En este caso, la diferencia más significativa es de una molécula de agua. Al encontrar la diferencia, se puede separar la combinación que usa el agua y resaltarla. Si esta diferencia menor se encontrara, por ejemplo, en un grupo fosfato (ácido fosfórico), también debería ser relativamente sencillo hallarla y separar el grupo. El supuesto es que la diferencia más pequeña debe expresar un grupo químico que distingue el sustrato del producto. Además, es de esperar que la tendencia entre los grupos que se desprenden de otros compuestos sean más pequeños y que un producto sea más grande que los sustratos en el anabolismo, lo cual también obedece a los supuestos de la regla de balance.

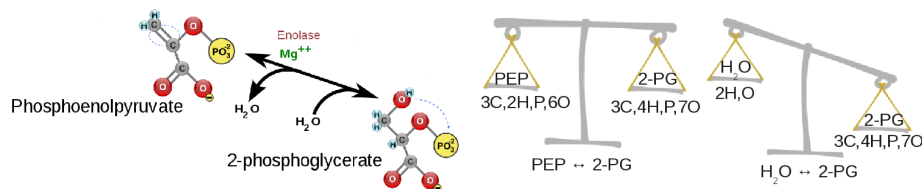


Figura 4.1: **Esquema general de la regla de balance.** Izquierda: reacción catalizada por la enolasa (R00658). Derecha: Esquema de las comparaciones de masa de las combinaciones disponibles. Las diferencias de masa entre los compuestos son el criterio principal que se usa para diferenciar y seleccionar.

La regla de balance considera que los pares sustrato–producto mejor resueltos reflejan la asociación más cercana entre compuestos, y que, en consecuencia, pares de compuestos con una mayor diferencia se parecen menos. En la figura 4.1, se puede discernir claramente que el fosfoenolpiruvato (PEP) está más cercanamente asociado al 2-fosfoglicerato (2-PG). El agua funge en esta reacción como un grupo saliente, que deja la ruta con la formación de

PEP. Puede argumentarse contra este acomodo que incluir pares de compuestos contruidos como en RPAIR/RCLASS, en los cuales un sustrato puede relacionarse con más de un producto de la misma reacción, representaría las interacciones existentes con mayor precisión que una partición que no repite un compuesto una vez que se usa. Esto funciona en el caso de RPAIR, porque el objetivo central de este conjunto de datos es exponer todas las relaciones válidas entre compuestos como pares. Si se intenta colocar en el contexto de una intensidad de asociación, el ejemplo muestra que no es trivial: queda el problema de definir cómo es que el agua se asocia más o menos con PEP o con 2-PG. La regla de balance evita este problema considerando al agua como asociada no a PEP o a 2-PG, sino al par $2\text{-PG} \leftrightarrow \text{PEP}$. El agua es, entonces, *explícitamente* un grupo saliente resultado de un evento químico que involucra 2-PG y PEP, lo cual se apega a la representación usual de esta reacción. Esta decisión de diseño enfoca al método a producir un ordenamiento rígido, lineal y fácil de entender y evaluar.

4.2.2. Regla de conteo

En ocasiones, la regla de balance no es suficiente para obtener un par sustrato-producto que pueda declararse como el par principal. Un ejemplo de la idea es la ATP:dGDP fosfotransferasa (R02090), que se ilustra en la figura 4.2. Este tipo de reacción entre nucleótidos se usa como paso para la transcripción. Si se tienen compuestos con cambios idénticos en toda la reacción, la regla de balance puede fallar. A pesar de que ambos compuestos pudieran agruparse, es difícil establecer qué par es más particular a la reacción, puesto que ninguno de los pares atraviesa un cambio mayor. De igual forma, si los compuestos son lo suficientemente parecidos, la regla de balance puede separar los pares que comparten el grupo que se transfiere en vez de los que involucran a la transición; en el ejemplo, podría ser un par $\text{ATP} \rightarrow \text{dGTP}$ en vez del par $\text{ATP} \rightarrow \text{ADP}$.

Para ayudar a resolver este problema, se recurrió a una medida que pudiera evaluar la relevancia relativa para un compuesto dentro de una reacción tomando en cuenta su abundancia en la red metabólica. Ésta regla es llamada de *conteo*, y toma el número de veces que una transformación dada se repite en todas las reacciones en el conjunto de datos. El supuesto tomado para la regla de conteo es que los compuestos que no son el objeto principal de transformación, como las coenzimas, se distribuyen de forma más amplia en rutas poco relacionadas entre sí. Por ende, se toma el supuesto de que los compuestos que más definen una reacción son los más únicos: los que se repiten menos a lo largo de todo el metabolismo, y por ende tienen una probabilidad mayor de ser exclusivos para esa reacción. Con base en este

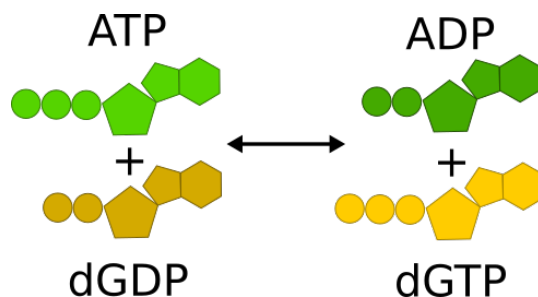


Figura 4.2: **Ejemplo de fallo para la regla de balance**, que plantea la necesidad de la regla de conteo. Los compuestos que son muy similares y sufren cambios idénticos son difíciles de distinguir en términos de masa. Un ejemplo es esta reacción entre ATP y dGDP.

critério, se puede proceder por eliminación de una forma similar a como se procede con la regla de balance, seleccionando las transformaciones que se repiten más.

La idea general de esta regla se ilustra en la figura 4.3. Por cada reacción, se van reuniendo las transformaciones de compuestos existentes, y se van agregando a una cuenta global. Posteriormente, cuando se busca aplicar la regla a una reacción, simplemente se busca el valor apropiado en la lista generada.

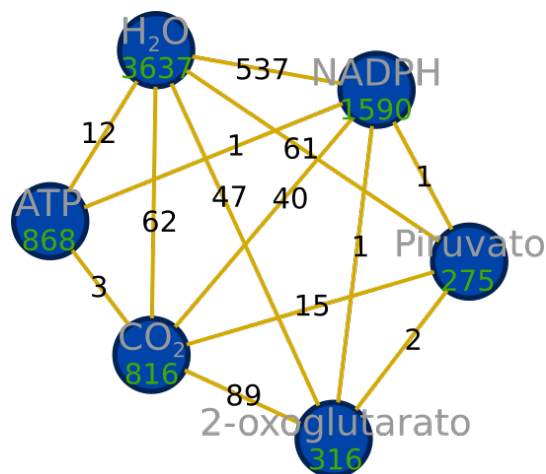


Figura 4.3: **Idea general de la regla de conteo**. Si se toman todas las interacciones posibles dentro de la red metabólica, es más coherente tomar las interacciones entre compuestos (resaltadas en amarillo) según su frecuencia en las reacciones del conjunto de datos que los compuestos mismos (en azul), puesto que es una medida más cercana al evento químico que se desea representar.

Como sugiere la figura 4.3, este concepto es similar al planteado por los métodos de topología del metabolismo, que buscan medir la asociación de compuestos en la red por su grado. La idea de la regla de conteo no es tomar los compuestos en sí, sino su ocurrencia conjunta en las reacciones. Tomar la ocurrencia de un par de compuestos, *juntos*, en las reacciones de la red, considerando que cada uno debe estar en un lado de la ecuación, refleja mejor la idea de la transformación de un par sustrato–producto, que es de importancia central al describir el metabolismo. Esta perspectiva retoma la idea de la asociación global de compuestos de una forma novedosa que busca apearse más a las características del sistema que se busca estudiar.

4.3. Implementación de las reglas

Uno de los objetivos de este proyecto fue implementar computacionalmente las reglas descritas en la sección anterior, para poder usarlas sobre cada reacción en la red metabólica seleccionada. Esta implementación, como se explicará a continuación, permitió establecer la cercanía de los compuestos para cada reacción, así como estructuras que nos permitieron clasificarlas. Más aún, la implementación también facilitó el análisis de los grupos de reacción derivados de dichas clasificaciones.

4.3.1. Asociación entre grupos de compuestos

Para generar los grupos de compuestos, se inició con la ecuación balanceada y se tomó cada lado de la ecuación por separado. Se calcularon todas las combinaciones de compuestos de cada lado separado y se parearon con todas las del otro. Esto es equivalente a decir que se tomó el producto Cartesiano de todas las combinaciones posibles de los compuestos de cada lado de la ecuación que tuvieran al menos un compuesto; esto es, eliminando pares que pudieran tener como elemento único en un lado de la ecuación el conjunto vacío. Por esta razón, a cada uno de los pares de combinaciones resultantes se le llamó Elemento del producto Cartesiano (ECP) (figura 4.4).

La generación de los ECPs se hizo como se describe a continuación (ver definiciones en Glosario). Sea c un compuesto cualquiera y C el conjunto de todos los compuestos en la red metabólica. Dado que en una reacción los compuestos pueden ser sustratos o productos, se puede definir un lado l de una reacción como

$$l = \{c_1, c_2 \dots c_q \in C\}; n(l) \geq 1 \quad (4.1)$$

siendo $n(l)$ la cardinalidad de l .

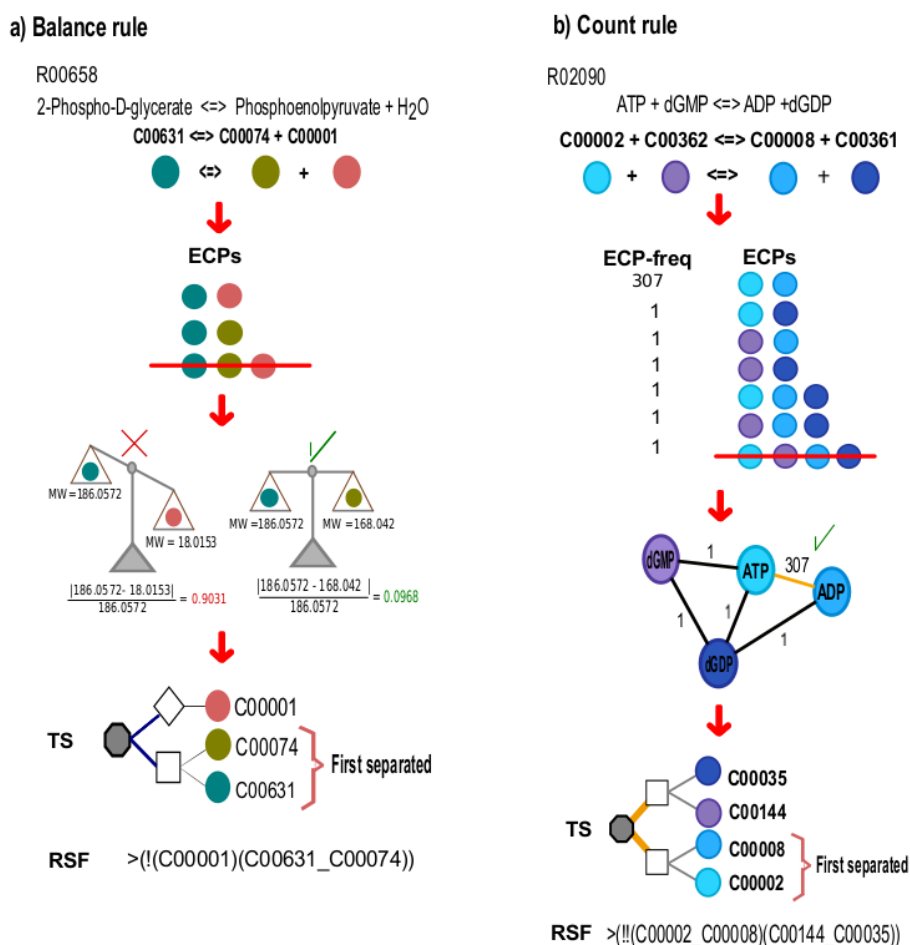


Figura 4.4: Caricatura que describe el algoritmo implementado para las reglas de balance y conteo. a) Regla de balance. Se usa como ejemplo la reacción R00658 de KEGG, mediada por la enolasa en la glicólisis. El balance se basa en la similitud de masas entre los dos lados de la ecuación en un ECP (*elemento del producto Cartesiano*; ver texto). Los ECPs cuyos lados sean más similares son seleccionados; la idea, como se ilustra, es unir los compuestos más parecidos a nivel de masa. b) Regla de conteo. Se usa como ejemplo la reacción R02090 de KEGG. El conteo se basa en la frecuencia de un ECP dado. En vez de tomar el número de conexiones de los nodos (círculos azules), se cuenta el número de reacciones en las cuales aparece una combinación de compuestos (líneas amarillas). En ambos casos, las combinaciones pueden ser de cualquier tamaño mientras no sean el total del ECP original.

Cabe recordar que KEGG, dentro del conjunto REACTION, no da un

significado específico a los lados de cada reacción, pudiendo ser cualquiera de ellos el de los sustratos o el de los productos (sección 1.2). Debido a esto, los lados no se pueden definir como *sustratos* o *productos*, sino solo de forma general, como *izquierdo* o *derecho*; esto permite a los ECPs manejar reacciones sin tener que forzar una direccionalidad. Esto también tiene otros efectos favorables, como se discute en la sección 4.4.

Con esto, si l_i es el lado izquierdo de una reacción cualquiera r y l_d su lado derecho, puede definirse r como un par ordenado:

$$r = (l_i, l_d) \quad (4.2)$$

Con esto puede expresarse el concepto formal de ECP. Siendo $\mathcal{P}(A)$ el conjunto potencia de un conjunto cualquiera A , se puede definir de forma general $L(l)$, la combinatoria de un lado l , como

$$L(l) = \{q \in \mathcal{P}(l) \mid n(q) \geq 1\} \quad (4.3)$$

Con esta ecuación, se define X_r , el producto Cartesiano para la reacción r , como

$$X_r = L(l_i) \times L(l_d) \quad (4.4)$$

El ECP x_r se define entonces como un elemento de este conjunto:

$$x_r = \{q \in X_r \mid q \neq r\} \quad (4.5)$$

Cabe subrayar que se ignora el ECP que representa el total de la reacción, puesto que no agrega información útil para el proceso; la razón se discute en la sección 4.4.

La ventaja principal del concepto de ECP es que da una forma de usar las reglas con grupos de compuestos de cualquier tamaño; la regla de balance puede aplicarse sobre los lados de un ECP (figura 4.4a) y la regla de conteo puede ser la cuenta de un ECP en el total de las reacciones (figura 4.4b). Esto hace posible usar las reglas de forma *exhaustiva*, tomándolas para formar posibles pares sustrato–producto cubriendo todas las combinaciones de compuestos. El poder explorar todas las posibilidades de combinación dentro de una reacción permite usar las reglas sin la necesidad de recurrir a las estructuras de los compuestos, a fin de evaluar que tan eficaces son para formar pares de compuestos viables sin tener que complementarlas con otros parámetros.

4.3.2. Implementación algorítmica de las reglas

Para poder usar las reglas, es necesario describir a detalle su forma final como un proceso, para exponer las ideas que involucran la aplicación del concepto. A continuación se hace esta descripción para ambas reglas.

Para aplicar la regla de balance sobre ECPs de tamaño arbitrario, se le ajusta a porcentajes; lo que se busca no son las transformaciones de menor masa, sino las transformaciones que implican el menor cambio de los compuestos involucrados expresadas como una tasa de cambio. Esto se hace siguiendo el supuesto de similitud que se espera de un par sustrato-producto. En el ejemplo de la figura 4.1, se espera que PEP se desprenda de 2-PG; en reacciones que usan ATP para obtener PO_4 , se espera que de éste se desprenda ADP; en otras reacciones se espera lo mismo del par NAD-NADH. No obstante, si se toma el ejemplo de la reacción R00756 (sección 4.1), la diferencia entre ambos pares es únicamente el grupo PO_4 , lo cual haría fallar a la regla de balance incluso si los compuestos son claramente diferentes a nivel de masa. Sin usar las estructuras, es difícil calcular qué relación tiene un compuesto con los otros usando la diferencia bruta, puesto que ésta puede coincidir entre pares diferentes, como en el ejemplo. Se hace necesario ver qué fracción es un compuesto de otro, y en general se espera que esta diferencia sea la menor posible, puesto que el evento esperado es que el compuesto sea donador o aceptor de un grupo, que es una fracción relativamente pequeña.

Con esto, la regla de balance se define a partir de $S[l(x)]$, la suma de masas de un lado l del ECP x , incluyendo los coeficientes de cada compuesto (Figura 4.4a). Siendo c un compuesto cualquiera, m_c el peso molecular de c y $q_{c,l(x)}$ el coeficiente para c en l y x :

$$S[l(x)] = \sum_{a=1}^{n(l)} q_{c,l(x)} m_c \quad (4.6)$$

Con base en estas sumas, $B(x)$, el resultado de la regla de balance para el ECP x , es el valor absoluto de la diferencia de estas sumas dividido entre la suma mayor:

$$B(x) = \frac{|S[l_i(x)] - S[l_d(x)]|}{\max(S[l_i(x)], S[l_d(x)])} \quad (4.7)$$

Tomar el valor absoluto en vez de la diferencia bruta compensa el hecho de que la base de datos no toma en cuenta una direccionalidad fija para las reacciones. La división de la ecuación 4.7 pretende generalizar el concepto de similitud de compuestos; como se empieza con la reacción completa, se procede a examinar qué compuestos conservan mejor la integridad de su

composición sin asumir *a priori* cuales están más relacionados, colocando la comparación en el contexto de la reacción particular. Cabe mencionar aquí que se prefiere el valor absoluto al cuadrado porque lo que se quiere ver son las proporciones; usar un cuadrado daría mayor peso a los valores más grandes, pero distorsionaría la representación de las proporciones de masa que se desea obtener.

En el caso de la regla de conteo, la implementación es trivial; simplemente consiste en contar el número de veces que un ECP aparece en todas las reacciones disponibles, tomando en cuenta los productos Cartesianos completos ($X(r)$) para cada reacción (figura 4.4b).

4.4. Aplicación recursiva

Para reducir los ECPs, se usó un método de dos estrategias base, la división y la recursión; el proceso se expresa en la figura 4.5. Desde la reacción completa, se genera el producto Cartesiano y se aplican las reglas. La regla de balance se aplica primero siempre, puesto que es una medida de asociación de compuestos que solo depende de la reacción actual. Si la regla de balance no puede reducir el conjunto de ECPs a una única posibilidad, entonces se aplica la regla de conteo sobre los ECPs que queden, con miras a reducir este conjunto a un ECP único. Si la regla de conteo tampoco es suficiente, la reacción se declara como imposible de manejar.

El método es divisivo porque usa las reglas para dividir la reacción original y así determinar qué compuestos están más asociados y agruparlos. Si se obtiene un ECP único aplicando las reglas como se estableció arriba, se selecciona y se separa. Tomando la reacción o el ECP que se esté dividiendo, se eliminan todos los compuestos que estén el ECP seleccionado; este nuevo conjunto se toma aparte, incluso si se trata de un solo compuesto aislado. Esto genera únicamente dos grupos de compuestos, cada uno de los cuales indica más cercanía para los compuestos que lo conforman dentro del grupo que entre grupos. Esta decisión de diseño se tomó porque si el grupo final es de un par de compuestos es más probable que ese par sea un par sustrato-producto, lo cual es la representación que plantean las reglas y el objetivo de crear los ECPs. Esto es visible sobre todo en el contexto de reacciones como la R00756, que sólo necesitan una división para generar pares sustrato-producto.

Como consecuencia de esto, puede discutirse por qué se ignora la reacción total; en una reacción bien curada la diferencia de masas entre ambos lados de la ecuación es cero. Por ende, dada la naturaleza divisiva del método, la reacción total resulta poco informativa al usar la regla de balance.

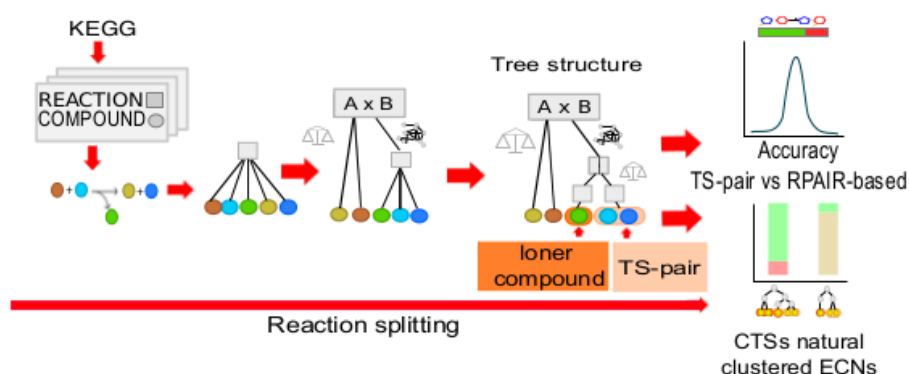


Figura 4.5: **Resumen gráfico del enfoque.** El enfoque toma la reacción completa de la base de datos (izquierda) y subdivide, dejando las cesiones de grupos implícitas en las diferencias entre los compuestos ligados a cada nodo. Éstos se van separando a cada paso (flecha roja *Reaction splitting*), generando una estructura de árbol para la reacción, ligando los compuestos como pares (*TS-pair*) o compuestos aislados (*loner compound*). Con estos árboles se hacen pruebas para validar (derecha): una prueba estadística basada en la coincidencia de los pares en los árboles con los de RPAIR (arriba) y una bioquímica comparando su coincidencia con los números EC (abajo).

Si un ECP tiene al menos un compuesto de cada lado y al menos dos compuestos en al menos un lado, las reglas vuelven a aplicarse sobre él, tomando las nuevas divisiones como procedentes del ECP original (figura 4.5). Cada uno de estos conjuntos es revisado para ver si es un par (un compuesto del lado izquierdo y uno del lado derecho) o un compuesto aislado. Si lo es, se deja como está; en caso contrario, se divide de la misma forma como derivado del ECP del que procede, conservando éste último como el paso de división anterior. Esta forma de aplicar el método minimiza los parámetros adicionales que las reglas requieren, puesto que el proceso se centra en aplicar las reglas que ya fueron definidas. No obstante, al volver a aplicar las reglas sobre las nuevas elecciones de ECP, a cada paso se genera información nueva sobre la distancia relativa entre los compuestos; cada compuesto tiene distintas distancias de los compuestos que fueron separados antes y de los que serán separados después. A esta estrategia se le llama *recursiva*; la repetición del proceso se basa en ampliar la información de pasos anteriores (figura 4.6). Al repetir el proceso divisivo básico, la estructura de división de la reacción se vuelve más detallada sin necesidad de agregar parámetros nuevos.

Esta forma de usar las reglas implica el problema de manejar interacciones de tres compuestos, como las de las lisis y ligaciones de compuestos, que

recursivo	5!	iterativo
fact(5)		fact(5)
5*fact(4)		1*1 = 1
5*4*fact(3)		1*2 = 2
5*4*3*fact(2)		2*3 = 6
5*4*3*2*fact(1)		6*4 = 24
5*4*3*2*1 = 120		24*5 = 120

Figura 4.6: **Concepto de recursión**, tomando como ejemplo la función factorial. La recursión (izquierda) toma un estado inicial y lo repite una y otra vez, reservando la información de pasos anteriores para calcular el estado final. Se distingue de la iteración (derecha) porque en la recursión cada paso modifica la cantidad de información que maneja el proceso, mientras que en la iteración la cantidad de información permanece constante.

forman parte de los ECPs que se deben examinar. Esto se manejó agregando el concepto de *compuesto aislado*; un compuesto que no se asocia directamente a otro compuesto, sino a un par de compuestos. Esto permite expresar un compuesto entrante o saliente como asociado al par que forma el resto de la molécula de cada lado del ECP (figura 4.5). Un ejemplo es la reacción catalizada por la enolasa (figura 4.1), que con una sola división genera el par 2-PG→PEP pero deja sola una molécula de agua. Esta molécula procede de la lisis de 2-PG, pero no es claramente separable del par; es producto de la transformación que el par implica. Por esto, es mejor expresarla como asociada al par: tanto PEP como H₂O son productos de la reacción, pero la estructura del par permite subrayar que PEP es más parecido al compuesto original.

El proceso recursivo continúa hasta tener ya sea pares o compuestos aislados en los estados finales o hasta que no se pueda dividir, creando la estructura que se toma como resultado para esa reacción. Es natural representar estructuras de este tipo como árboles, puesto que cada paso de división permanece marcado para detallar el historial de la parte divisiva; el camino que siguió el proceso hasta cada par o compuesto aislado. En consecuencia, a la estructura final que se calcula para cada reacción se le llama *estructura de árbol* o *tree structure* (TS).

La TS, por la estructura recursiva de la división, preserva toda la información sobre las divisiones sucesivas; esto revela el proceso que originó

cada par final de compuestos, como se representa en la figura 4.5. Los pares obtenidos son los mejores candidatos a ser pares sustrato–producto de acuerdo con las reglas, lo cual permite examinarlos como tales. La presencia de los compuestos aislados permite señalar qué compuestos tienen un rol de grupos liberados de un par sustrato–producto; esto debería dejar los pares con mayor evidencia. Además, como los pares resultantes son pares sustrato–producto, la direccionalidad no tiene un peso determinante en la estructura de la TS; ambos compuestos quedan conectados al mismo nodo intermedio, y los compuestos aislados siguen siendo los únicos conectados a su nodo.

Las decisiones de diseño del método se enfocan a usar las reglas suponiendo que efectivamente generan estructuras familiares (pares sustrato–producto). A partir de esto, se pueden examinar las propiedades de las TSs para encontrar una forma de examinarlas con base en este supuesto.

Parte III

Resultados y Discusión

Capítulo 5

Generación de las clases

5.1. Agrupamiento de las TSs

El objetivo de usar las TSs de las reacciones es representar cómo cada par sustrato–producto se acomoda en cada reacción; los compuestos aislados representarían entonces pasos de transición similares a los grupos de tres presentes en lisis de compuestos. En este paso, sin embargo, cabe preguntarse cuánto se pueden parecer las TSs entre sí; después de todo, existen muchas reacciones que son casi iguales, y que únicamente usan coenzimas levemente diferentes. La pregunta es válida en este punto porque si se obtienen estas similitudes para reacciones parecidas se puede confirmar que el método es coherente para reacciones con mecanismos esencialmente idénticos, una propiedad que se espera si el método efectivamente maneja propiedades de reacciones. La figura 5.1 presenta un ejemplo de esto. Las reacciones de los casos 5.1a y 5.1b son casi idénticas; la única diferencia es que una usa NAD y otra usa NADP.

Al observar otras reacciones puede notarse que, aunque el patrón de compuestos pueda no ser igual, la reacción puede compartir su TS básica con otras reacciones. Tal es el caso en la reacción de la figura 5.1c; la reacción usa compuestos principales diferentes, pero la TS sigue siendo la misma. Esto sugiere que las TSs pueden implicar un patrón más general, independiente de los compuestos que se transformen en una reacción dada. Para examinar la posibilidad de este patrón, todas las reacciones que compartan su estructura de TS — entendida como la topología del árbol para la TS y las reglas que fueron usadas para cada paso específico — son tomadas como un solo grupo de TSs. A este tipo de grupo se le llama *cluster de estructuras de árbol* o *cluster of tree structures* (CTS). Todos los CTSs obtenidos se muestran en la tabla del apéndice A.1.

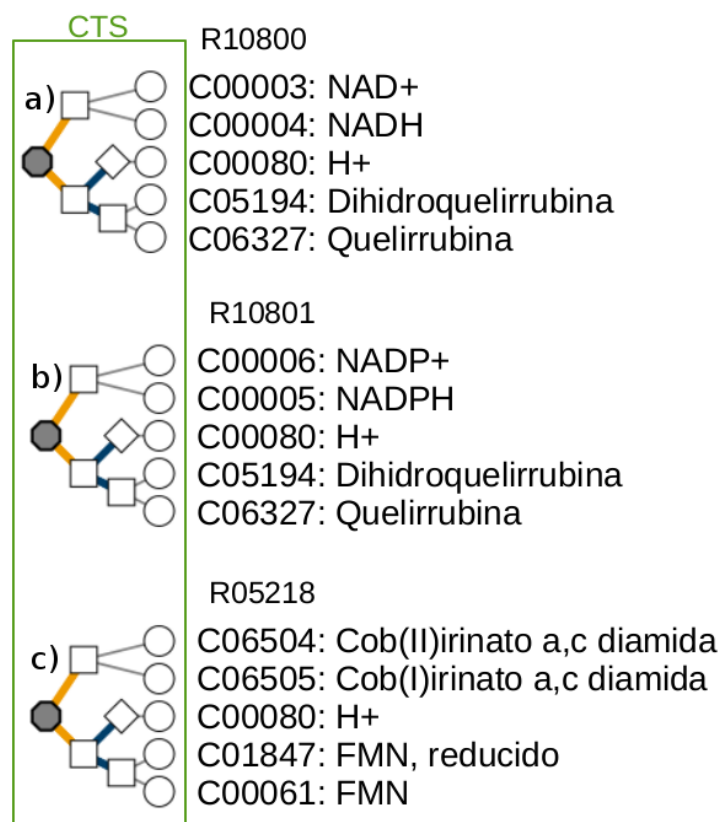



Figura 5.1: **Ciertas TSs se parecen independientemente de los compuestos que usan.** Las reacciones en las partes a) y b) tienen compuestos muy similares, por lo cual es casi trivial suponer que las TSs deben parecerse. Sin embargo, la parte c) usa compuestos diferentes y pese a ello su TS se parece a las de las partes a) y b).

Las TSs y los CTSs se representaron de tres formas: un árbol como se espera de forma gráfica; por formato JSON (*JavaScript Object Notation*), adecuado para manejo automático; y las que se llamaron representaciones *de texto y compacta*, para manejar TSs y CTSs como texto plano.

La tabla 5.1 ilustra las representaciones de texto y compacta para la reacción R00760. Cabe subrayar las diferencias entre el formato de texto y el compacto: “>” en el compacto es “*root*” en el de texto, y “!” en el compacto es “*balance*” en el de texto. El formato compacto también usa “!!!” para las subdivisiones por conteo, que el de texto representa explícitamente como “*count*”. Las representaciones de texto se leen de izquierda a derecha, lo que corresponde a observar la representación gráfica de arriba a abajo.

Tabla 5.1: **Representaciones en diferentes formatos para TSs y CTSs.** Se toma la reacción R00760 como ejemplo. Para el formato gráfico: Nodos: gris, raíz; círculo, compuesto; cuadrado, par. Aristas: línea azul, separado por balance; línea delgada, liga nodo/compuesto.

Representación	Ejemplo
Árbol, texto	root(balance(C00095_C00085)(C00002_C00008))
Árbol, compacto	>(! (C00095_C00085)(C00002_C00008))
CTS, compacto	>!(C_C)(C_C)
CTS, gráfico	

5.2. Características generales de las TSs

Se tomaron los datos de la base de datos KEGG en su versión 2015 [22], de los conjuntos COMPOUND, RPAIR, RCLASS y REACTION. De COMPOUND se tomaron los pesos moleculares de los compuestos; de RPAIR, los pares que están en el conjunto RCLASS junto con su categoría en RPAIR; y de REACTION, las fórmulas de las reacciones y los RPairs que les corresponden. Se excluyeron reacciones con compuestos sin pesos moleculares en COMPOUND y aquellas con compuestos con coeficientes y secciones de fórmula variables.

5.3. Descripción de los datos utilizados para la generación de las TSs

De las 9910 reacciones en REACTION para el set de 2015, se tomaron 7526 reacciones con pesos moleculares definidos y coeficientes no variables para todos los compuestos en COMPOUND y REACTION. De éstas, 1099 se ignoraron por ser de solo un par de compuestos y 35 por no poder separarse en pares o aislados. Esto se hizo para realizar el análisis sobre un conjunto de reacciones bien descritas cuya solución no fuera obvia. Por ejemplo, una reacción que consiste sólo en un par de compuestos se devuelve como tal sin

realizar trabajo y las reacciones que dan conjuntos de más de dos compuestos no están equilibradas para realizar un análisis a nivel de masas, por lo cual no hay información suficiente para explicar aciertos o errores del método en estas reacciones.

5.4. Consideraciones sobre los agrupamientos

El proceso de remoción de las reacciones con curación no óptima dejó 6392 reacciones con una TS. Agruparlas por topología generó los 71 CTSs que se presentan en la tabla A.1. De estos CTSs, se seleccionaron aquellos con al menos 10 reacciones cada uno, teniendo en cuenta que un número de reacciones demasiado reducido en un CTS no permitiría hacer pruebas estadísticas confiables. Por ello, al final preservamos para análisis posteriores a 22 CTSs para realizar pruebas enfocadas a detectar propiedades de los CTSs en vez de propiedades del conjunto completo (6279 reacciones; 98.23 %).

Cabe subrayar que si un par de TSs fueron idénticas a nivel de topología pero diferentes en sus pasos de resolución, como en los casos de los CTSs 1 y 19 y los CTSs 8 y 22 (figura 5.2), se consideraron CTSs distintos. Esto se debe a que las propiedades que separan los ECPs en un paso dado son diferentes; la regla de balance en un caso y la regla de conteo en el otro. Incluso cuando la estructura matemática de la TS sea igual en ambos casos, ambos tipos de reacción tienen propiedades de intercambio químico diferentes que llevan a esa separación, indicando en particular que la regla de balance no es suficiente para separar los compuestos. La necesidad de incluir la regla de conteo, como se plantea en la sección 4.2, surge de la presencia de transiciones de grupo idénticas entre ECPs muy similares, lo cual involucra interacciones que no son claramente distinguibles únicamente por masas.

Definir estas características generales hace posible usarlas para hacer una validación formal del método: determinar qué tan viable es el método viendo en qué medida se ajusta a las propiedades bioquímicas que busca representar. La parte siguiente se enfoca en desarrollar esta validación y mostrar el ajuste que tiene el método con las propiedades bioquímicas de los pares sustrato-producto, que son la base en la cual se sustenta.

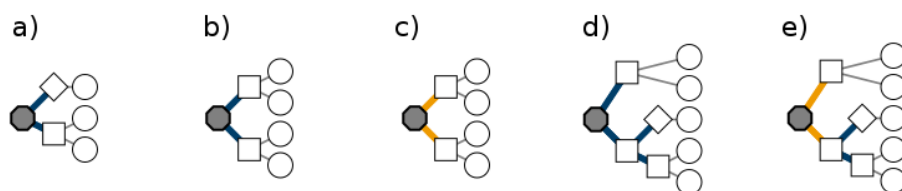


Figura 5.2: **Los CTSs representan configuraciones únicas.** De izquierda a derecha: a) CTS 2, una estructura de un par y un compuesto aislado resuelta por balance; b) CTS 1, una estructura de dos pares de compuestos resuelta por balance; c) CTS 19, una estructura de dos pares de compuestos resuelta por conteo; d) CTS 8, una estructura compleja resuelta por balance; e) CTS 22, una estructura compleja resuelta por ambas reglas. Nodos: gris, raíz; círculo, compuesto; cuadrado, par; rombo, saliente. Aristas: línea azul, separado por balance; línea naranja, separado por conteo; línea delgada, liga nodo/compuesto.

Capítulo 6

Exploración de las TSs

6.1. Exploración estadística del método

6.1.1. Análisis estadístico de pares

Los patrones que representan los CTSs sugieren que puede haber una relación entre la TS y las transformaciones de compuestos en una reacción química. Esta correspondencia se evaluó a dos niveles. El primero se enfoca en la precisión de las reglas empleadas y el segundo en la precisión de los grupos generados. Esta sección se enfoca en la revisión de la efectividad de las reglas y su implementación, y la siguiente en confirmar la precisión de los grupos.

Para determinar la efectividad de las reglas, se debe recurrir al supuesto del cual se desprenden: las reglas elegidas son propiedades de pares sustrato-producto del metabolismo, y por ello estos pares pueden detectarse con su uso. Para determinar si este supuesto se cumple, se debe comparar el conjunto de pares que se generan con las TSs y los pares sustrato-producto curados en una base de datos. Para este fin, se eligió el conjunto de datos RCLASS, complementándolo con datos en RPAIR para la versión de 2015 de KEGG, disponible en nuestro grupo. Este conjunto de datos contiene estos pares curados correspondientes a las reacciones del metabolismo.

Es necesario establecer un método exacto que pueda detectar correctamente los pares, adaptándose a las características de cada conjunto de datos para asegurar que sean comparables. Como ya se explicó, los pares en las TSs son comparables con los pares de RPAIR [26], pero es necesario determinar una medida de acierto para precisar la similitud entre ambos conjuntos de pares debido a la presencia de los compuestos aislados; hay *RPairs* que podrían no recuperarse porque un compuesto en un *RPair* queda separado de su correspondiente en la TS. Un ejemplo es la reacción R00025 (etilnitronato \rightarrow

acetaldehído + nitrito), dentro de la cual el etilnitronato (C18091) está en los *RPairs* C00084_C18091 (con acetaldehído) y C00088_C18091 (con nitrito) usando FMN (C01847_C00061). La TS despliega el *RPair* C00088_C18091, pero no C00084_C18091, porque el acetaldehído da un valor de balance de menor parecido, lo cual favorece que se incluya aparte del etilnitronato.

Debido a que la coincidencia de los pares no es directa, se optó por tomar a RPAIR/RCLASS como estándar principal (*gold standard* en el *argot* de la estadística), para evaluar solamente la forma en que los pares de las TSs aparecen en RPAIR/RCLASS asumiendo que este conjunto de datos posee una tasa menor de errores debido a que está revisado manualmente, incluso aunque es posible encontrar fallas de curación en KEGG [29].

Para poder hacer la evaluación de los pares de las TSs, es necesario tomar en cuenta que éstos pueden o no aparecer en RPAIR/RCLASS, y además el hecho de que RCLASS ha sido actualizado para incluir más pares. Por ende, también se requiere poder estimar el cambio de los datos al paso del tiempo. Para lograrlo, se prefirió un enfoque bayesiano, que haga posible considerar un estimado del cambio de los pares de KEGG. La prueba elegida fue simplemente tomar los datos, tomar la proporción de coincidencias y no coincidencias y evaluar las proporciones con una distribución bayesiana; el hecho de que el método elegido usa una combinatoria hace imposible evaluarlo de forma combinatoria por definición.

La distribución usada se eligió tomando en cuenta que tomar un par de TS cualquiera y determinar si es o no un *RPair* para esa misma reacción es un ensayo de Bernoulli, puesto que un par de TS únicamente puede o ser o no ser un *RPair* registrado. Por ello, se tomó la distribución Beta, puesto que es la previa conjugada de la Bernoulli; si se toman datos con una distribución de Bernoulli, la distribución Beta estima las probabilidades bayesianas para un posible espacio futuro de datos. Se tomó la media de la distribución como medida de precisión de pares, puesto que es el valor estimado de la probabilidad de éxito del ensayo de Bernoulli.

Para todo el conjunto de 6392 reacciones curadas, la precisión fue de 81.48 %, como se muestra en la figura 6.1a. Tomando esta precisión, se puede apreciar que, dada la simpleza de las reglas y del proceso empleado para aplicarlas, el método es sumamente preciso en general, lo cual respalda la elección de reglas realizada por la búsqueda en la literatura.

Luego de haber confirmado la precisión general del método, la Beta fue evaluada independientemente para todos los pares dentro de cada CTS con al menos 10 reacciones; como es de esperar, los CTSs dan distintos niveles de precisión. Para determinar la confianza obtenida para cada CTS, ésta se definió en tres niveles, según el valor de la media de la Beta: alta ($\bar{x} \geq 0.8$), media ($0.8 > \bar{x} \geq 0.6$) y baja ($0.6 > \bar{x}$).

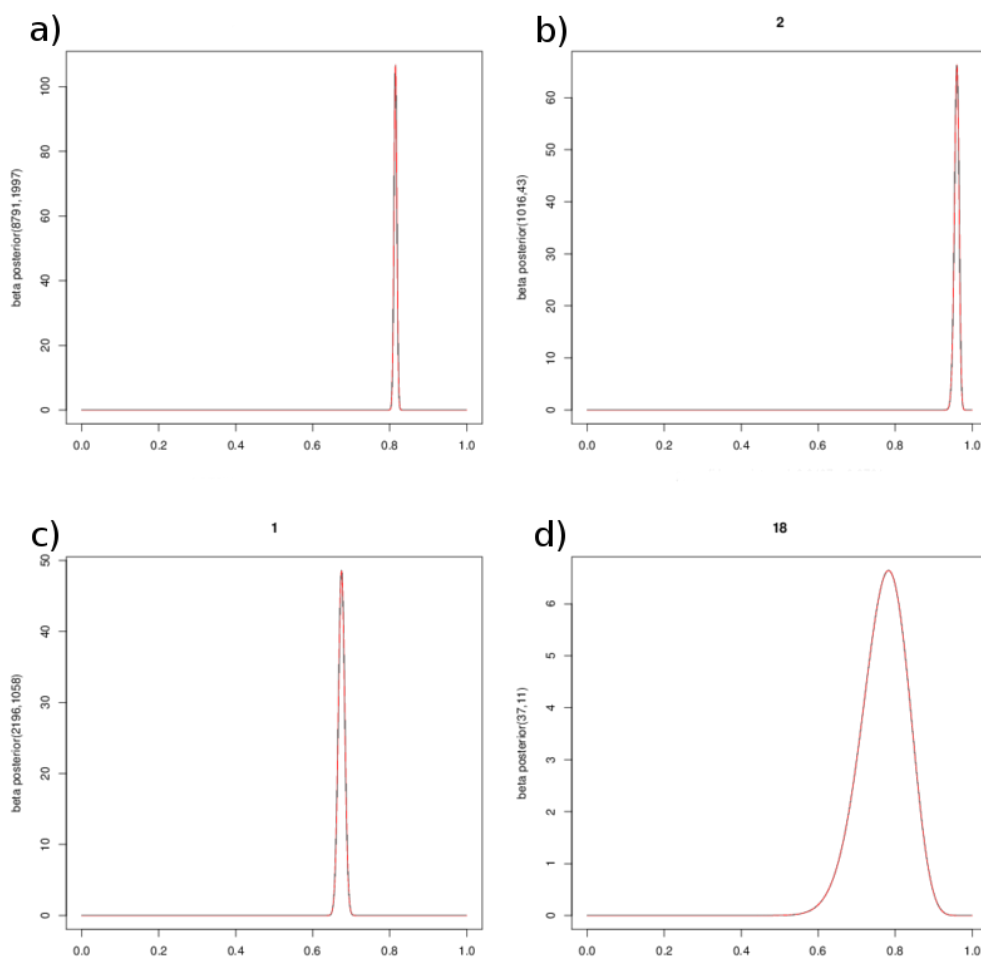


Figura 6.1: **La precisión del presente enfoque varía entre CTSs.** Ejemplos de las curvas de distribución posterior para θ obtenidas con la distribución Beta (ver texto), con cada conjunto de parámetros (aciertos,errores). a) Total del conjunto de datos; b) CTS 2, una distribución de precisión alta; c) CTS 1, de precisión media; d) CTS 18, de precisión media, pero pocos datos.

Al aplicar los resultados de la prueba estadística sobre los CTSs, se apreció que fue alta para 8 CTSs (3879 reacciones), media para 10 (2198 reacciones) y baja para 4 (202 reacciones). La tabla 6.1 da los valores para los 22 CTSs usados para esta prueba; la figura 6.1 presenta ejemplos. El CTS 2, por ejemplo, es de alta precisión (figura 6.1b): 1016 pares de 1059 concuerdan, y la amplitud de la distribución es mínima por el gran número de reacciones en el CTS.

Tabla 6.1: Datos de distribución Beta para el conjunto de TSs.

	Pares totales	Pares acertados	Pares fallidos	Precisión media
Total	10788	8791	1997	0.8148
CTS 1	3254	2196	1058	0.6748
CTS 2	1059	1016	43	0.9586
CTS 3	1028	926	102	0.9000
CTS 4	1794	1756	38	0.9783
CTS 5	878	873	5	0.9933
CTS 6	698	690	8	0.9872
CTS 7	450	206	244	0.4580
CTS 8	286	193	93	0.6737
CTS 9	206	151	55	0.7309
CTS 10	291	196	95	0.6724
CTS 11	246	163	83	0.6613
CTS 12	66	58	8	0.8680
CTS 13	108	86	22	0.7911
CTS 14	84	58	26	0.6862
CTS 15	58	57	1	0.9684
CTS 16	78	29	49	0.3749
CTS 17	72	48	24	0.6623
CTS 18	48	37	11	0.7602
CTS 19	28	13	15	0.4667
CTS 20	12	11	1	0.8564
CTS 21	24	13	11	0.5384
CTS 22	20	15	5	0.7264

A este nivel, es posible apreciar ciertos detalles que apuntan a un patrón en el acomodo de los pares en cada CTS. El CTS 18 (figura 6.1d), por ejemplo, es de precisión media: 37 pares de 48 son correctos. Como existen menos elementos en el CTS para medir, la amplitud de la curva crece. Una inspección cercana revela que estas reacciones son mediadas por ligasas (grupos de ECs 6.3.2.-, 6.3.5.-), dos de ellas por la tobramicina carbamoiltransferasa (EC 6.1.2.2).

Por otra parte, una excepción a la asociación entre precisión alta y poca amplitud es el CTS 1 (figura 6.1c); de sus 3254 pares, solo 2196 son concordantes, lo que le da una precisión media (67.48 %). Este CTS se compone de

reacciones de diversos ECs, sobre todo transferasas (grupo 2), pero también de otros grupos de enzimas, como oxidorreductasas (grupos 1.1.5.-, 1.21.3.-) e hidrolasas (grupo 3.5.4.-). Esto indica que una regla general puede no ser suficiente para algunas reacciones, o que incluso el grupo puede contener más de un grupo de ECs, lo cual se discute más adelante. Pese a esto, los resultados señalan que hay ECs que predominan en los distintos CTSs, por lo que puede haber un patrón de asociación del acomodo de los pares entre los CTSs y transformaciones químicas reportadas en la base de datos. Es necesario entonces buscar un modo de validar esta asociación en general.

6.1.2. Análisis de los CTSs mediante clases de pares sustrato–producto

Los resultados arriba descritos sugieren que los CTSs contienen reacciones con transiciones químicas similares. Para examinar esta relación, se decidió examinar si la tendencia de los pares obtenidos es efectivamente coincidente con pares sustrato–producto reportados.

Para examinar las tendencias de los pares de cada TS, éstos fueron clasificados por el nivel de precisión evaluado con la Beta dentro de su CTS (alta, media o baja), discutido en la sección 6.1.1. Para asignar la clasificación química de los pares, se usó la propuesta por Kotera et. al. para los *RPairs* [26], discutida en la sección 1.2 (*main*, *cofac*, *ligase*, *trans*, *leave*).

Para relacionar ambas medidas, se agregó una clasificación de reacciones por su proporción de pares encontrados y no encontrados en RCLASS y se agruparon los pares individuales sobre este esquema. Las reacciones fueron clasificadas como *acertadas* (todos los pares están en RCLASS), *fallidas* (ningún par está en RCLASS), o mixtas (hay al menos un par en RCLASS y un par que no está en RCLASS). Cada par en cada reacción fue entonces clasificado, según fuera el caso, como un par acertado en una reacción acertada (*entirely predicted pair*, EPP), un par acertado en una reacción mixta (*mixed positive pair*, MPP), un par fallido en una reacción mixta (*mixed failed pair*, MFP) o un par fallido en una reacción fallida (*failed pair*, FP).

La figura 6.2 despliega la frecuencia de reacciones que tiene cada categoría bajo este esquema triple. De los 10,788 pares disponibles en las 6,392 reacciones usadas, 969 no se encontraron ni en RPAIR ni en RCLASS, dejando 9,819 pares. De éstos (7,672; 78.13 %) caen dentro de la categoría EPP. Dentro de EPP, la mayoría de los pares se concentran en los grupos *main* (5,757, 58.63 % del total) y *cofac* (1,839, 18.73 % del total), con alta precisión (*main*, 3,728 de 5,757; *cofac*, 1,620 de 1,839). Este resultado, por sí mismo, muestra que las reglas son eficaces para encontrar pares que representan coenzimas

de oxidorreducciones y, más importantemente, pares sustrato–producto que se consideran pares principales. Este hecho respalda la eficacia de las reglas en la labor de encontrar pares para las categorías más abundantes, algo que no deja de sorprender dada la simpleza del proceso global. Cabe señalar que EPP también tiene uno de los grupos *ligase* más grandes, a pesar de que sea un miembro minoritario de la categoría (68 de 7,672, confianza media).

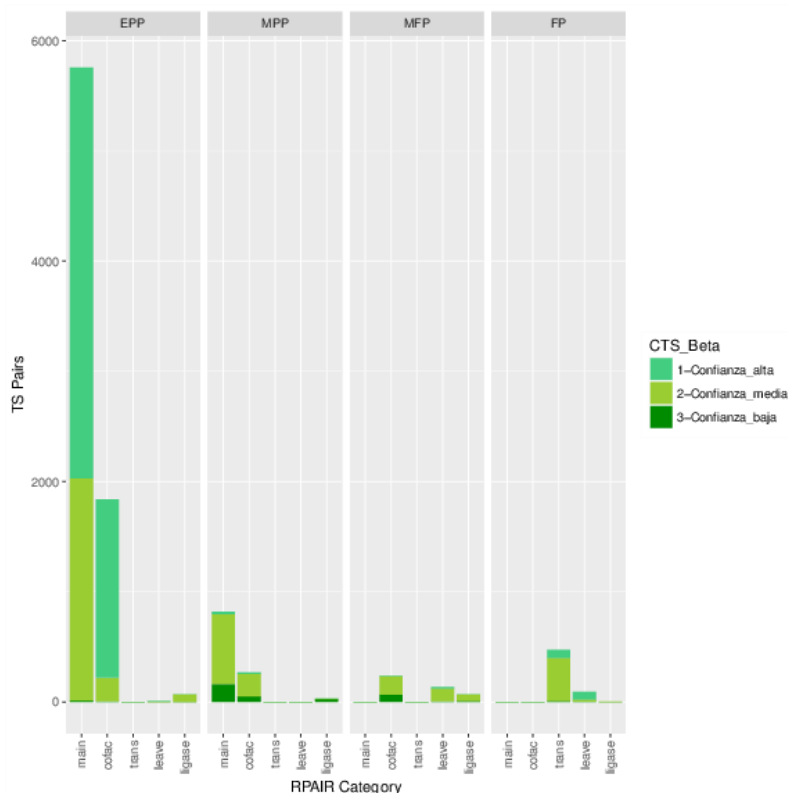


Figura 6.2: La precisión es mayor para las clases *main* y *cofac* de Kotera et. al. Abundancia de pares según clases tomadas de datos de RPAIR: la clasificación de Kotera et. al. (columnas repetidas; *main*, *cofac*, *trans*, *leave*, *ligase*), la precisión del par según la media de la Beta en el análisis bayesiano y un esquema de proporción de pares acertados por reacción. La correlación bayesiana se indica por color de las barras (confianza alta, media y baja). El esquema de pares acertados se señala en cada columna única: EPP (pares en reacciones completamente acertadas), MPP (pares positivos en reacciones mixtas), MFP (pares negativos en reacciones mixtas) y FP (pares en reacciones completamente fallidas).

Las reacciones mixtas, en general, tienden a caer dentro de la precisión media. La categoría MPP, por ejemplo, es mucho más pequeña que la EPP (de 9,819, 1,119; 11.40 %), pero también concentra grupos *main* (818 de 1119)

y *cofac* (268 de 1119) de precisión media (*main*, 634 de 818; *cofac*, 204 de 268). La categoría MFP, por su parte (de 9,819, 452; 4.60 %), concentra varios pares *cofac* (238), un grupo *ligase* mayoritario (68) y el grupo *leave* más grande (136). En esta categoría, predomina la precisión media (*cofac*, 168 de 238; *ligase*, 55 de 68; *leave*, 117 de 136). Este resultado conecta a las reacciones mixtas con la clase *leave*, así como la categoría EPP queda conectada con las clases *main* y *cofac*.

La categoría FP (de 9,819, 576; 5.87 %) representa las diferencias más marcadas del método al clasificar pares. Esta categoría sobresale por tener el grupo *trans* más grande (474 de 576) con precisión media (391 de 474) y por concentrar pares incorrectos del CTS 1 (precisión media; 383 *trans*, 11 *leave*). Esto conecta a la categoría FP con la clase *trans*, pero también mayoritariamente con el CTS 1, una observación notable dado que las demás clases comprenden varios CTSs. Al examinar los tipos de reacción dentro de la categoría, se observa que son en su mayoría glicosiltransferasas (de hexosas y pentosas), fosfotransferasas con aceptor alcohol, y transferasas de grupos arilo y alquilo alternos al metilo. Un artículo relacionado con KEGG reporta que los métodos usados para alimentar la base de datos tienen problemas para trabajar en particular con glicotransferasas [35], lo cual indica que incluso las discordancias que se obtienen corresponden a los grupos donde la base de datos tiene conjuntos de anotaciones con evidencia más débil. Esto da a entender las bases químicas que explican la coherencia de los pares obtenidos de las TSs con los existentes en KEGG.

Determinar la precisión general del método para encontrar pares sustrato-producto es importante debido a que estas transiciones son las más usadas para determinar rutas metabólicas. Faust et al. [9] compararon el desempeño de usar buscadores de rutas que combinan listas curadas de pares de compuestos con otros parámetros, mostrando que el mejor enfoque para trazar rutas combina la lista curada de RPAIR con un esquema de pesado que penaliza conexiones con compuestos frecuentes. Los resultados de esta sección indican que este método da un esquema de trazado de interacciones internas que aparta preferencialmente estos compuestos frecuentes del resto de la reacción sin necesidad de hacer una penalización explícita, puesto que ésta queda ya comprendida dentro de las reglas de balance y conteo.

Los resultados apuntan a esta conclusión; la mayoría de los pares encontrados caen en árboles de alto nivel de confianza, y la mayoría de éstos, a su vez, cae dentro de la categoría *main* de la clasificación de Kotera. La clasificación obtenida y sus proporciones concuerdan con las que produce un proceso de curación manual, lo cual reafirma la eficacia general de las reglas en que se sustenta el enfoque. También hace posible ubicar las debilidades presentes; los pares *leave* y *trans* se resuelven con menor eficacia y caen en

clases de reacciones precisas, como es el caso de las glicosiltransferasas en la categoría FP de pares. De igual forma, la clase *ligase* queda distribuida entre dos categorías de precisión de pares, lo cual hace la posibilidad de un nexo menos clara que en las demás categorías. Esto hace de los pares *ligase*, *leave* y *trans* candidatos para curación manual con miras a mejorar el método, posiblemente usando alguna regla adicional que comprenda las características de estas transiciones.

Estos resultados, en su conjunto, respaldan la presencia de una tendencia de los CTS a concentrar reacciones en categorías que reflejan transiciones químicas similares. Con base en esta observación, cabe inspeccionar si los CTSs correlacionan con grupos de enzimas de funciones similares.

Capítulo 7

Exploración de los CTSs

7.1. Los CTSs tienden a agrupar categorías enzimáticas

Dado que las clases de Kotera sugieren que los CTSs agrupan transiciones químicas similares es relacionarlos con categorías enzimáticas establecidas. Tomar una clasificación enzimática que refleje el tipo de catálisis que se lleva a cabo en las distintas reacciones de un CTS permitiría, en un futuro, evaluar si existen relaciones evolutivas entre las enzimas de estos grupos.

Con este fin, se recurrió a la clasificación de la Comisión de Enzimas, expresada por los ECs. Los ECs se eligieron porque, a pesar de ser una clasificación de enzimas, están diseñados para agruparlas según características de catálisis, lo cual ayuda a traslapar las reacciones con una categoría de catálisis y una de enzimas. KEGG ya registra las asociaciones entre reacción y enzima por sus ECs, por lo cual estos datos pueden aprovecharse para este análisis. Se agruparon las reacciones por su CTS y por su EC, agrupando los ECs a niveles de 1, 2 y 3 dígitos; esto permite detectar posibles patrones a diferentes niveles de detalle según los ECs. El objetivo es observar la relación de un CTS con una categoría de ECs particular, y no el hacer comparaciones funcionales entre categorías de ECs.

Para poder apreciar un patrón, primero se observa el conjunto de datos para visualizar alguna tendencia que pueda ser explorada estadísticamente. Con este fin, se agruparon las anotaciones de ECs por reacción y se hizo una cuenta simple de estas asociaciones por CTS. Se contaron todos los puntos disponibles para tomar en cuenta las asociaciones múltiples que pueden darse entre enzimas y reacciones que están presentes en KEGG. La figura 7.1 presenta la acumulación de reacciones por CTS para grupos de ECs a primer dígito. Los CTSs están ordenados por su abundancia de reacciones, pero las

barras no corresponden a este orden por la posibilidad de que una reacción esté asociada a más de un EC.

A pesar de que puede haber varios grupos de ECs dentro de un solo CTS, se nota una fuerte tendencia de los CTSs a concentrar un único primer dígito de EC; el CTS 1 concentra transferasas, el 2 liasas, el 3 hidrolasas y el 4 oxidorreducciones. Esta tendencia se aprecia en el resto de la distribución.

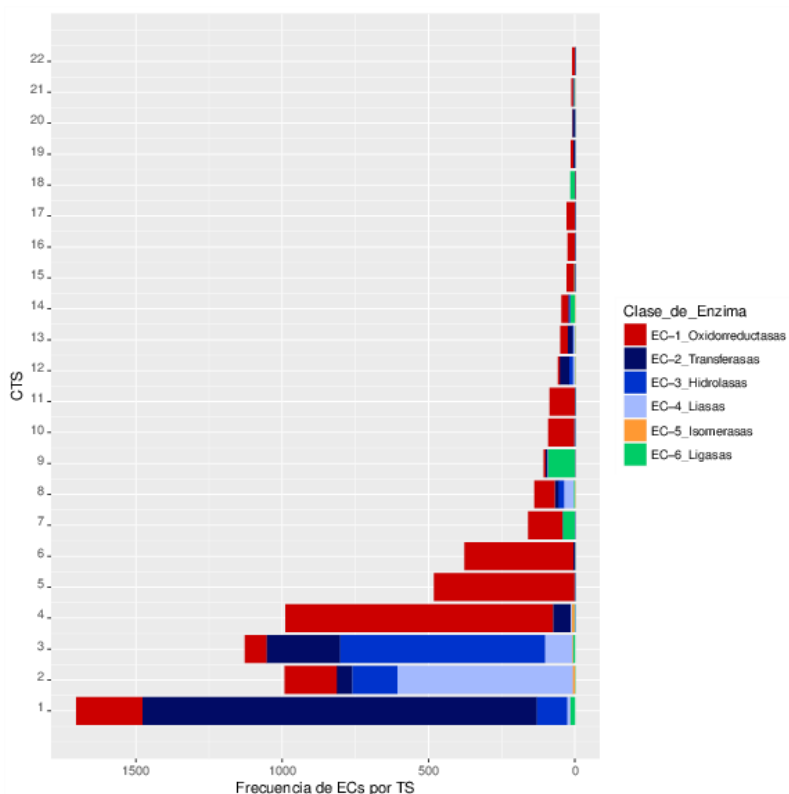


Figura 7.1: **Los CTSs concentran diferentes grupos enzimáticos generales.** Gráfica de frecuencia de reacciones, contrastando CTSs contra grupos de ECs a primer dígito. Abcisa: Frecuencia de reacciones; Ordenada, CTSs; Color, grupos de ECs, según la leyenda.

Se observa en la figura 7.1 que la mayoría de los CTSs tienden a concentrar el grupo de ECs 1, de oxidorreductasas. Esta tendencia es consistente con la idea de que el funcionamiento del metabolismo es primariamente oxidativo, pero hace pensar que el grupo 1 no tiene una asociación clara con un CTS particular. Por el contrario, los grupos 2 (transferasas; CTS 1), 3 (hidrolasas; CTS 3), 4 (liasas; CTS 2) y 6 (ligasas; CTS 9) están muy concentrados en un solo CTS; incluso cuando hay grupos más pequeños que parecen signifi-

cativos, el grueso de las anotaciones de EC para las reacciones caen dentro de un solo CTS.

Cabe mencionar que las isomerasas (grupo EC 5) casi no se notan en la figura 7.1; los grupos están en los CTSs 2, 3 y 4. Esto se debe a que, al eliminar las reacciones que solo consistían en un par de compuestos antes de generar los árboles, también fueron eliminadas casi todas las isomerasas en el set original. Pese a esto, las isomerasas restantes se concentran en los CTSs mencionados sin aparecer en ningún otro, lo cual es coherente con la tendencia observada.

Estos resultados apuntan a un acomodo preferencial de las anotaciones de EC en CTSs particulares. Aunque no es sustituto de una prueba estadística formal, sí señalan tendencias que pueden explorarse con este tipo de pruebas. Una prueba estadística confirmaría estas tendencias y daría información sobre las tendencias que no pueden apreciarse con una cuenta simple; además, aclararía el caso de las oxidorreductasas, que queda ambiguo a este nivel.

Para validar la tendencia de la figura 7.1, es necesario medir la significancia de la correlación entre CTSs y grupos de ECs. Esto daría mayor peso a las tendencias más amplias de la figura 7.1 y permitiría apreciar detalles sobre las oxidorreductasas. Además, ayudaría a visualizar qué tan asociado está cierto grupo de ECs con un CTS, algo que no es fácil de ver con una cuenta simple debido a las proporciones dadas por las diferencias de tamaño entre los CTSs.

Usando los mismos datos que en la prueba anterior, se midió el enriquecimiento de grupos de ECs para cada CTS con una prueba exacta de Fisher. Esto es porque, en el nivel en el cual pueden compararse — la correspondencia mutua de anotaciones de CTS e EC sobre una reacción — ambos tipos de dato pueden contarse como categóricos, y lo que se quiere ver es el nivel de asociación entre ambos tipos. La prueba de Fisher está diseñada para hacer este tipo de comparaciones entre categorías múltiples. En este caso, la hipótesis nula contra la cual compara la prueba de Fisher es la independencia de ambos tipos de dato, que puede interpretarse como una distribución uniforme con pequeñas variaciones, que corresponde a un acomodo al azar.

La prueba de Fisher puede dar dos tipos de valor para la asociación entre un par de tipos de dato a comparar: el *p-value* y el *odds-ratio*. El *p-value*, o valor *p*, da la probabilidad de que la asociación que se observa sea igual a la que se propone por la hipótesis nula; es una medida de significancia respecto del azar. El *odds-ratio*, o tasa de posibilidades, da la probabilidad de que la asociación que se observa sea igual al nivel usual en el que se presenta uno de los tipos de dato en toda la población; es una medida de significancia respecto del mismo conjunto de datos, o *enriquecimiento*. El *p-value* ayuda a determinar que la asociación que se detecte sea no explicable

por el azar, y el *odds-ratio* ayuda a determinar que la asociación que se detecte sea efectivamente una asociación justificable dadas las proporciones del conjunto de datos.

Para poder detectar la asociación y determinar la significancia de los patrones detectados, se agruparon las reacciones de los CTSs de al menos 10 reacciones por CTS y por EC. Se comparó la frecuencia de cada grupo de ECs a primer, segundo y tercer dígito entre CTSs, por las mismas razones que en la sección anterior. Se calculó la prueba exacta de Fisher para el total de este conjunto de datos. Se ajustaron los *p-values* por medio de la corrección de Benjamini–Hochberg para reducir la susceptibilidad a falsos positivos [4] y se desecharon las correlaciones con un *p-value* de más de 0.05 (la probabilidad de que la asociación sea explicable por el azar es de más de 5%). De las restantes, se tomó el *odd-ratio* como medida de enriquecimiento, puesto que este número da la asociación relativa que puede compararse entre los valores significativos.

La figura 7.2 muestra el resultado la prueba de Fisher a un dígito. Las correlaciones más altas se encuentran entre los CTSs con pocas reacciones; 16, 17 y 22 en el grupo 1 (redox). Esto se debe a que todas las reacciones en esos CTSs están dentro de su grupo de ECs, lo cual impide a la prueba comparar alternativas; el valor de enriquecimiento escapa a la precisión del dispositivo de cómputo. En estos casos, la gráfica ilustra un nivel de enriquecimiento solo para los límites de logaritmo base 10 del *odds-ratio* que van de -3 a 3; estos límites dan valores de hasta 3 órdenes de magnitud del enriquecimiento, y la mayoría de los valores más significativos caen dentro de este rango. Los CTSs 20 y 21 no están igual de enriquecidos, puesto que incluyen unas pocas reacciones fuera del grupo de ECs mayoritario; el tamaño del grupo cambia el enriquecimiento ajustándolo a las proporciones.

En cuanto a los CTSs restantes, se observa que 11 de ellos están preferencialmente enriquecidos solamente en un primer dígito de EC con un $\log(\text{Odd-ratio})$ mayor a 0 (más de 1 vez respecto del común). Estos CTSs dan una asociación significativa incluso tomando en cuenta que el EC ocupa un rango de valores muy amplio. No obstante, estas asociaciones pueden explicarse mejor si se toman los valores de respecto de las categorías de EC a tercer dígito (tabla A.2). Los CTSs tienden a enriquecerse más significativamente en cierto tipo de mecanismo incluso dentro del grupo de oxidorreductasas. Algunos casos son muy marcados: el CTS 22 concentra reacciones del grupo 1.16.-.-, que oxidan iones metálicos usando principalmente a O_2 y derivados de flavina como aceptores; el CTS 6 concentra el grupo 1.17.1.-, que oxida grupos CH o CH_2 y usa a NAD o NADP como aceptor; y el CTS 17 concentra el grupo 1.4.1.-, que actúa en grupos HC-NH_2 y usa a NAD o NADP como aceptor.

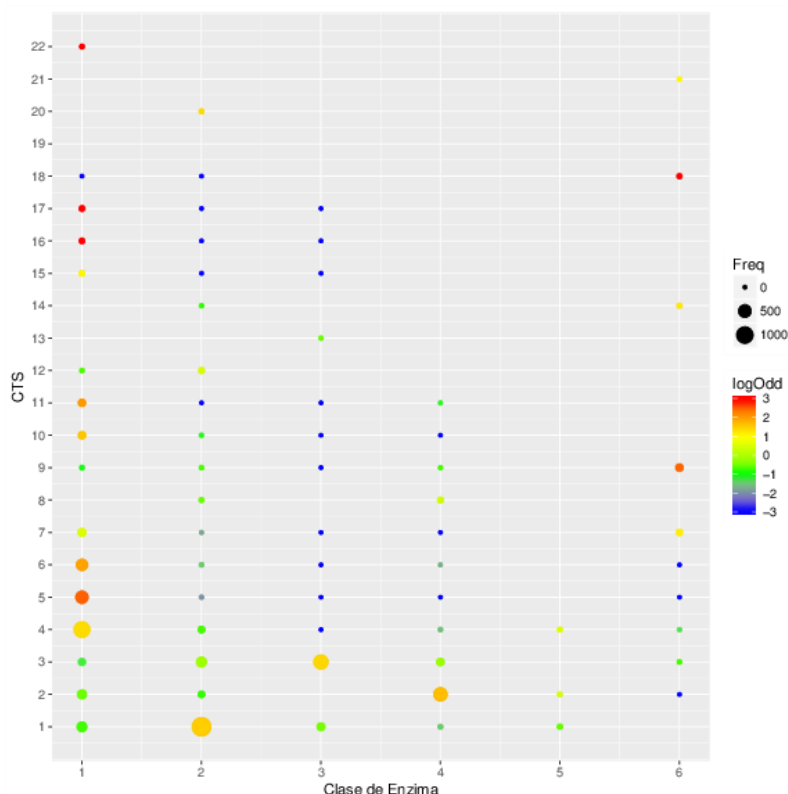


Figura 7.2: **Los CTSs están enriquecidos selectivamente en diferentes grupos de ECs.** Abcisa: CTSs; Ordenada, grupos de ECs; Tamaño de punto, frecuencia de reacciones; Color, $\log(\text{Odd-ratio})$, según la leyenda. Los valores se limitan a $\log(\text{Odd-ratio})$ de 3 a -3; los valores que van más allá se incluyen, pero su valor no se precisa más allá de los límites.

Un caso que permite ver el grado en el que se especializan los CTSs es el grupo de ECs 1.14.-.-, que actúa en donadores pareados asistido por O_2 . Este grupo está difundido entre varios CTSs, pero el enriquecimiento se concentra en un matiz particular o marca una mayor asociación para otro grupo. Por ejemplo, el grupo 1.14.13.-, que usa a NAD o NADP como donador e incorpora un átomo de oxígeno a la molécula, está más asociado al CTS 5. El grupo 1.14.-.- también está en los CTSs 11 y 15, pero el CTS 15 está más enriquecido en el grupo 1.14.16.-, que también incorpora un átomo de oxígeno pero usa tetrahydrobiopterina como donador (transformando aminoácidos aromáticos como fenilalanina o triptofano y compuestos parecidos como el antranilato), y el CTS 11 está levemente más enriquecido en el grupo 1.13.12.-, que también incorpora un átomo de oxígeno pero actúa sobre donadores sencillos, que también son compuestos con anillos purínicos (teo-

bromina: R07963-4; cafeína: R07971-2). El factor común entre los tres CTSs es la incorporación de oxígeno a la molécula y la formación de agua, pero los mecanismos más asociados son diferentes, y estas diferencias se reflejan en los cambios de topología de cada CTS. Las reacciones con donadores sencillos liberan un grupo que tiende a ser el formaldehído, mientras que las reacciones con donadores pareados no liberan ningún grupo adicional.

Este resultado está relacionado con una característica importante de la clasificación de los ECs. Las categorías de segundo y tercer dígito de ECs no son normalmente comparables. Lo son dentro de la clase de ECs 1, como se muestra en la tabla 7.1, pero no en las demás. Esto refleja el hecho de que las oxidorreductasas tienden a emplear mecanismos generales que son muy parecidos, mientras que los otros tipos de enzima son más variables en sus catálisis. Este hecho incrementa el alcance de la interpretación que puede hacerse del resultado arriba descrito; si hubiera una similitud a nivel de las coenzimas que se usan entre las coincidencias de los CTSs 5 y 11, ésta sería visiblemente detectable y haría más discutible la interpretación. Como esta similitud de las coenzimas no se da, cada grupo es más separable y puede argumentarse con más confianza una diferencia en el mecanismo.

Esta especialización en la separación de los ECs también ayuda a separar otros grupos de CTSs. Se observa, por ejemplo, que los CTSs 2, 4 y 7 están enriquecidos con un $\log(Odd-ratio)$ mayor a 0 en dos grupos de ECs. El segundo grupo enriquecido en los CTSs 2 (liasas) y 4 (redox) es el de los ECs 5 (isomerasas). Una inspección cercana reveló asociaciones de interés; para el CTS 4, todas las reacciones en el grupo 5 son Δ -isomerasas de esteroides. Para el CTS 2, las reacciones del grupo 5 se ubicaron en el EC 5.5.1.17 (cloromuconato cicloisomerasa; 7 reacciones) y el EC 5.5.1.11, que también transforma cloromuconato; ambos ECs liberan un grupo HCl . Las dos reacciones en el EC 5.5.1.17 que quedaron fuera del CTS 2 no liberan el grupo halógeno de la molécula, lo cual indica que, para este caso, el CTS refleja un tipo de catálisis particular para este EC.

Kotera et al. [26] y Rahman et al. [39] han desarrollado métodos para predecir pares de compuestos basados en ubicar centros de reacción, que permiten agrupar reacciones por su EC con ayuda de métodos de clustering previamente elaborados e independientes. La prueba de Fisher confirma que el presente enfoque agrupa de forma *natural*, sin ayuda de métodos adicionales, los patrones de uso de compuestos de las reacciones en las clases enzimáticas de los ECs. Esta observación es de gran importancia, puesto que muestra que este método es completo con sólo el uso recursivo de las reglas de balance y conteo; sólo con estas dos reglas es posible capturar propiedades que pueden conformar una clasificación independiente de reacciones.

La prueba de Fisher también ayuda a explicar las discrepancias encontra-

Tabla 7.1: **Sólo los ECs de las oxidorreductasas son mutuamente comparables.** Ejemplos de desgloses de ECs en enzimas individuales.

EC	Primer dígito	Segundo dígito	Tercer dígito	Cuarto dígito
1.1.5.9	Oxidorreductasas	Actúa en grupo CH–OH de donantes	Aceptor quinona o similar	glucosa deshidrogenasa
1.1.1.1	Oxidorreductasas	Actúa en grupo CH–OH de donantes	Aceptor NAD(P)	alcohol deshidrogenasa
1.3.1.1	Oxidorreductasas	Actúa en grupo CH–CH de donantes	Aceptor NAD(P)	dihidropirimidina deshidrogenasa
2.3.2.1	Transferasas	Aciltransferasas	Aminoaciltransferasas	Péptido-glutamina transferasa
2.3.1.1	Transferasas	Aciltransferasas	Grupos no aminoacilo	Glutamato N-acetiltransferasa
2.4.2.1	Transferasas	Glicosiltransferasas	Pentosiltransferasas	Purina-nucleósido fosforilasa

das en el trabajo con la Beta — como la del CTS 1 — que a pesar de poderse resolver en gran cantidad, tienen una confianza menor explicable por la variabilidad en su contenido de clases de ECs. Este resultado no contradice la tendencia observada, puesto que muchas reacciones en KEGG están asociadas a más de un EC, algo explicable por la existencia de enzimas promiscuas y de dominios múltiples, que son más bien comunes en el metabolismo [24]. De igual forma, los datos de la prueba de Fisher para las categorías a tercer dígito explican esta tendencia a detalle; distintos CTSs concentran diferentes clases de oxidorreductasas. Además, dado que sus ECs son los más comparables, como se ilustra en la tabla 7.1, es posible apreciar las diferencias de catálisis entre CTSs, lo cual revela que cada topología efectivamente puede reflejar diferencias en el mecanismo de reacción.

Estos resultados confirman la existencia de un buen nivel de correlación entre CTSs e ECs, y por ende que el presente enfoque categoriza clases catalíticas de forma independiente, sin necesidad de parámetros adicionales.

Capítulo 8

Conclusiones y Perspectivas

8.1. Conclusiones

En este trabajo se presentó un método de clasificación de reacciones basado en la asociación de los compuestos como pares sustrato–producto. Debido a que se asume que las rutas metabólicas transforman un compuesto de forma gradual, la atención se centra en maximizar la precisión para obtener los pares principales y se reduce para la contribución de otros compuestos en una reacción. Este trabajo muestra que evaluar los diferentes roles de los compuestos de forma local puede contribuir de forma significativa a la precisión de una búsqueda de pares sustrato–producto, incluso cuando se hace por medio de reglas simples.

El presente enfoque además revela que los CTSs están asociados con especies enzimáticas documentadas en una base de datos curada a mano. Los CTSs además pueden indicar diferencias importantes entre los mecanismos de reacciones relacionadas. Esto se pone de manifiesto al examinar los bloques de las oxidorreductasas y de las isomerasas, que tienen propiedades ligeramente diferentes dependiendo del CTS al cual están asociados.

En su conjunto, el trabajo muestra que el rol de un compuesto en una reacción es informativo para clasificarla. Por ende, estos roles no pueden descartarse de forma automática; pueden ser aprovechados de esta forma para contribuir a la precisión en la búsqueda de rutas metabólicas con menos trabajo computacional.

8.2. Perspectivas

El presente enfoque tiene potencial como una herramienta valiosa en el trabajo con reacciones y rutas. La prueba basada en la clasificación de Ko-

tera et. al. revela que el enfoque puede encontrar pares sustrato-producto relevantes para las reacciones, lo cual puede hacerlo un complemento eficaz para un método de predicción de rutas. Si se comparan las TSs que se forman para las reacciones a lo largo de una ruta, podrían analizarse patrones de uso de CTSs en la ruta o detectarse propiedades basadas en los pares sustrato-producto obtenidos; aquellos con niveles de precisión altos darían un desempeño superior al de los pares no consistentes.

De igual forma, las categorías tomadas de la verificación basada en ECs se elaboraron únicamente sobre reacciones con un EC asignado. Sin embargo, el presente enfoque confirma el acomodo de las reacciones en los CTSs mediante la prueba de enriquecimiento, lo cual da al método poder predictivo potencial para reacciones sin un EC.

Los resultados presentes también dan una guía para determinar qué reacciones son candidatos para curación manual, puesto que las reacciones cuyos pares no coinciden con los de RPAIR/RCLASS usando este proceso también presentan dificultades para los procesos establecidos de la base de datos.

El trabajo posterior se enfocará en incrementar la eficacia del enfoque, agregando parámetros como reglas adicionales que mejoren la predicción de pares y transformaciones enzimáticas. Una posible regla se basaría en parámetros termodinámicos, que ayudarían a determinar la direccionalidad de una reacción sin la necesidad de establecerla *a priori*.



De igual forma, el agrupamiento que hacen los CTSs de reacciones con ECs similares sugiere que es razonable explorar una hipótesis evolutiva que explique el acomodo de los CTSs. Sin embargo, la aplicabilidad directa de KEGG en estudios de este tipo puede cuestionarse debido a la dependencia de KEGG de datos de homología, que no necesariamente pueden reflejar la historia de una enzima. Una opción para este tipo de trabajo es la base de datos MANET [25], que intenta contestar una pregunta similar usando datos de dominios de proteína. El presente método puede usarse con los datos de MANET para comparar las TSs de cada reacción con los dominios registrados como un punto de partida para un trabajo evolutivo.

Apéndice A

Datos suplementarios

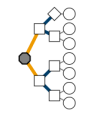
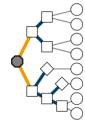





A.1. Clusters de estructuras de árbol (CTSs).

Los CTSs representan las estructuras básicas de las reacciones examinadas por el presente método, independientes de los compuestos específicos que usen. Manejar las reacciones de esta forma permite agruparlas naturalmente en estas estructuras, que se demuestra corresponden con categorías enzimáticas establecidas. La tabla de este apéndice muestra los 71 CTSs obtenidos. Sólo los primeros 22 se usaron en pruebas estadísticas, puesto que tienen al menos 10 reacciones. Para los grafos, el código de color es el mismo que el especificado en la sección 5.1. Para los nodos: gris, raíz; círculo, compuesto; cuadrado, par. Para las aristas: línea azul, separado por balance; línea naranja, separado por conteo; línea delgada, liga nodo/compuesto.

ID	Representación compacta	Reacciones	Grafo
1	$>(!(\text{C_C})(\text{C_C}))$	1627	
2	$>(!(\text{C})(\text{C_C}))$	1059	

ID	Representación compacta	Reacciones	Grafo
3	$>(!(\text{C})(!(\text{C})(\text{C_C})))$	1028	
4	$>(!(\text{C})(!(\text{C_C})(\text{C_C})))$	897	
5	$>(!(\text{C})(!(!(\text{C})(!(\text{C})(\text{C_C}))) (\text{C_C})))$	439	
6	$>(!(\text{C})(!(!(\text{C})(\text{C_C})) (\text{C_C})))$	349	
7	$>(!(\text{C_C})(!(\text{C_C})(\text{C_C})))$	150	
8	$>(!(\text{C_C})(!(\text{C})(\text{C_C})))$	143	
9	$>(!(!(\text{C})(\text{C_C})) (!(\text{C})(\text{C_C})))$	103	

ID	Representación compacta	Reacciones	Grafo
10	$>(!(\text{C})(!(\text{C_C})(\text{C_C}))(\text{C_C}))$	97	
11	$>(!(\text{C})(!(\text{C_C})(!(\text{C})(\text{C_C}))) (\text{C_C}))$	82	
12	$>(!(\text{C})(!(\text{C})(!(\text{C})(\text{C_C}))))$	66	
13	$>(!(\text{C})(\text{C_C}))(\text{C_C})$	54	
14	$>!(\text{C_C})(!(\text{C})(!(\text{C})(\text{C_C}))))$	42	
15	$>(!(\text{C})(!(\text{C})(\text{C_C})))(\text{C_C})$	29	
16	$>(!(\text{C_C})(!(\text{C})(\text{C_C})))(\text{C_C})$	26	

ID	Representación compacta	Reacciones	Grafo
17	$>(!(!(\text{C})(\text{C_C}))!(\text{C_C})(\text{C_C}))$	24	
18	$>(!(!(\text{C_C})(\text{C_C}))!(\text{C})(!(\text{C})(\text{C_C})))$	16	
19	$>(!(\text{C_C})(\text{C_C}))$	14	
20	$>!(\text{C})(!(\text{C})(!(\text{C})(!(\text{C})(\text{C_C}))))$	12	
21	$>!(\text{C})(!(\text{C_C})(!(\text{C})(\text{C_C})))$	12	
22	$>(!(\text{C_C})(!(\text{C})(\text{C_C}))$	10	
23	$>!(\text{C_C})(!(\text{C_C})(!(\text{C})(\text{C_C})))$	8	

ID	Representación compacta	Reacciones	Grafo
24	$>(!(\text{C_C})(!(\text{C})(\text{C_C}))(\text{C_C}))$	8	
25	$>(!(\text{C})(!(\text{C})(\text{C_C}))(\text{C_C}))(\text{C_C}))$	5	
26	$>(!(\text{C})(\text{C_C}))(!(\text{C_C})(\text{C_C}))$	5	
27	$>(!(\text{C})(\text{C_C}))(!(\text{C})(!(\text{C})(\text{C_C})))$	5	
28	$>(!(\text{C})(!(\text{C})(\text{C_C})))$	5	
29	$>(!(\text{C})(!(\text{C})(!(\text{C})(!(\text{C})(\text{C_C}))))(\text{C_C}))$	4	
30	$>(!(\text{C})(!(\text{C})(\text{C_C}))(!(\text{C})(\text{C_C})))$	4	

ID	Representación compacta	Reacciones	Grafo
31	$>(!!(C_C)(C_C))(C_C)$	4	
32	$>(!!(C)(C_C))(!(C)(C_C))$	4	
33	$>(!(C)!(!(C)!(C_C)!(C)!(C)(C_C))))(C_C))$	3	
34	$>(!(C)!(!(C)!(!(C)(C_C))(C_C)))(C_C))$	3	
35	$>(!!(C)!(C)(C_C))!(C)!(C)(C_C))$	3	
36	$>(!!(C_C)(C_C))(C_C)$	3	
37	$>(!!(C)(C_C))(C_C)$	3	

ID	Representación compacta	Reacciones	Grafo
38	$>(!(\text{C})(!(\text{C})(\text{C_C}))!(\text{C})(\text{C_C})(\text{C_C})))$	2	
39	$>(!(\text{C})(!(\text{C_C})(!(\text{C})(\text{C_C})(\text{C_C})))$	2	
40	$>(!(\text{C})(!(\text{C})(!(\text{C})(\text{C_C}))!(\text{C})(\text{C_C})))$	2	
41	$>(!(\text{C})(\text{C_C})(!(\text{C})(\text{C_C})(\text{C_C})))$	2	
42	$>(!(\text{C})(!(\text{C})(\text{C_C}))!(\text{C})(!(\text{C})(\text{C_C})))$	2	
43	$>(!(\text{C})(\text{C_C})(!(\text{C})(\text{C_C})(\text{C_C})))$	2	
44	$>(!(\text{C})(!(\text{C})(!(\text{C})(!(\text{C})(!(\text{C})(\text{C_C}))))))$	2	

ID	Representación compacta	Reacciones	Grafo
45	$>(!(\text{C_C})(!(\text{C})(!(\text{C_C})(\text{C_C}))))$	2	
46	$>(!(!(C)(!(C)(C_C)))(!(C)(C_C)))$	2	
47	$>(!(!(C_C)(C_C))(!(C)(C_C)))$	2	
48	$>(!(\text{C_C})(!(\text{C})(!(\text{C})(\text{C_C}))))$	2	
49	$>(!(!(C)(!(C)(C_C)))(C_C))$	2	
50	$>(!(!(C)(!(C)(C_C)))(!(C_C)(!(C)(C_C)(C_C))))$	1	
51	$>(!(\text{C})(!(\text{C})(!(\text{C})(\text{C_C}))(\text{C}(C_C))))(\text{C_C}))$	1	

ID	Representación compacta	Reacciones	Grafo
52	$>(!(!(\text{C})(!\text{C})(\text{C_C}))(!\text{C})(!\text{C})(!\text{C})(!\text{C})(\text{C_C}))))$	1	
53	$>(!(\text{C})(!(\text{C})(!\text{C})(!\text{C_C})(\text{C_C}))))(\text{C_C}))$	1	
54	$>(!(\text{C})(!(\text{C_C})(\text{C_C}))(!\text{C})(!\text{C})(\text{C_C}))))$	1	
55	$>!(\text{C_C})(!(\text{C})(!\text{C})(\text{C_C}))(!\text{C})(\text{C_C}))))$	1	
56	$>(!(!(\text{C})(!\text{C_C})(\text{C_C}))(!\text{C})(!\text{C})(\text{C_C}))))$	1	
57	$>(!(\text{C})(!(\text{C_C})(!\text{C})(!\text{C_C})(\text{C_C}))))$	1	
58	$>(!(\text{C})(!(\text{C})(!\text{C_C})(\text{C_C}))))(\text{C_C}))$	1	

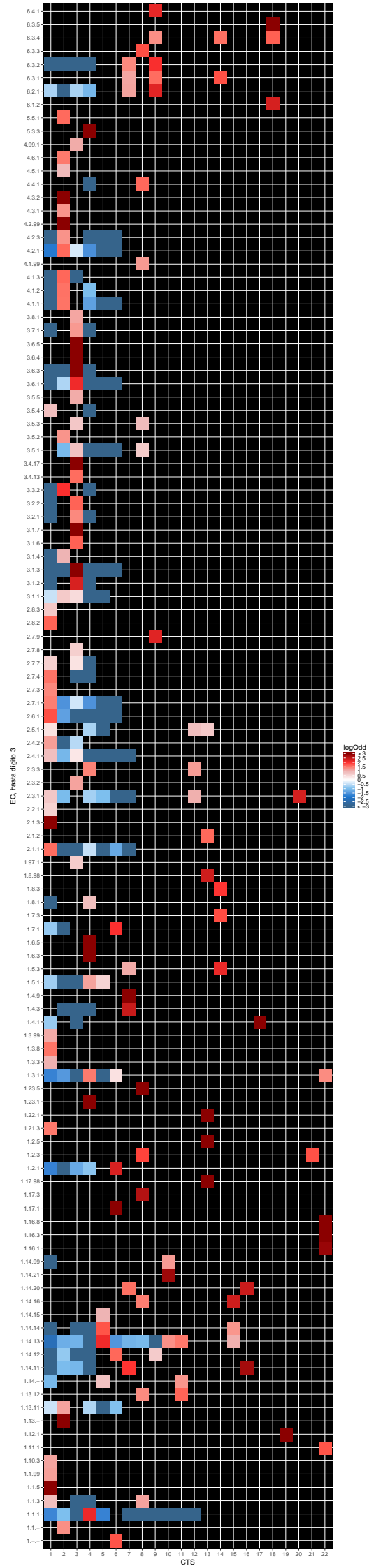
ID	Representación compacta	Reacciones	Grafo
59	$>(!(\text{C_C})(!(\text{C})(!(\text{C_C})(!(\text{C})(\text{C_C}))))))$	1	
60	$>(!(\text{C_C})(!(\text{C_C})(!(\text{C})(!(\text{C})(\text{C_C}))))))$	1	
61	$>(!(\text{C_C})(!(\text{C_C})(!(\text{C_C})(\text{C_C}))))$	1	
62	$>(!(!(\text{C})(\text{C_C}))(!(\text{C})(!(\text{C_C})(\text{C_C}))))$	1	
63	$>(!(\text{C_C})(!!(!(\text{C})(!(\text{C})(\text{C_C}))) (\text{C_C})))$	1	
64	$>(!(!(\text{C_C})(!(\text{C})(\text{C_C})))(!(\text{C})(\text{C_C})))$	1	
65	$>(!(!(\text{C})(\text{C_C}))(!(\text{C})(!(\text{C})(\text{C_C}))))$	1	

ID	Representación compacta	Reacciones	Grafo
66	$>(!!(C)!(C)!(C)(C_C)))(C_C)$	1	
67	$>(!!(C_C)!(C)(C_C))(C_C)$	1	
68	$>(!!(C_C)!(C)(C_C))(C_C)$	1	
69	$>(!!(C_C)!(C_C)(C_C))$	1	
70	$>!(C)!(C_C)(C_C)$	1	
71	$>!(C)(C_C)$	1	

A.2. Relación de CTSs con clases de ECs a tercer dígito.

Los CTSs representan clases generales de reacciones. Por ende, cabe preguntarse cómo encajan con clases de catálisis ya reportadas. Los ECs representan estas especies catalíticas como enzimas, con la ventaja de poder agruparlas en clases por sus características. Esta figura presenta los puntos de enriquecimiento más significativos para los CTSs respecto de clases de ECs a tercer dígito. Se omiten los enriquecimientos con un valor p mayor o igual a 0.05; al igual que en la sección 7.1, el color de la celda representa el $\log(Odd - ratio)$, según la leyenda. Los valores se limitan a $\log(Odd - ratio)$ de 3 a -3; los valores que van más allá se incluyen, pero su valor no se precisa fuera de los límites.

La figura se presenta en el archivo 2 a mayor tamaño para facilitar su revisión.



Glosario

árbol Topología en la cual a un nodo raíz se conectan varios nodos, y a cada uno de éstos pueden conectarse otros tantos, y así sucesivamente. Una clase notable es el *árbol binario*: a todos los nodos que reciben conexiones se conectan únicamente dos nodos. Ver *topología*. 33, 36

anabolismo Parte del metabolismo que usa bloques simples y energía para sintetizar macromoléculas. 1, 2, 23, 24

catabolismo Parte del metabolismo que degrada macromoléculas y compuestos energéticos para obtener energía y moléculas simples. 1, 2, 23

clustering Proporción de conexión de un nodo en un grafo respecto del máximo teórico posible para ese grafo. Si un nodo A se conecta con otros seis en un grafo de diez nodos, el clustering c de A se calcula por $c_A = 6 \div 10 = 0.6$. Ver *grado*, *grafo*. 10

conjunto En matemáticas, una colección de entes, que puede ser considerada un ente como tal. Los entes que conforman el conjunto, cuando se les refiere respecto del conjunto, son llamados *elementos* del conjunto. Por ejemplo, un conjunto que contiene los elementos a , b y c se representa como $\{a, b, c\}$. 27, 29

conjunto potencia El conjunto potencia $\mathcal{P}(A)$ de un conjunto A es el conjunto de todos los conjuntos que es posible construir a partir de todos los elementos de A , incluyendo al conjunto vacío y a A mismo. Ver *conjunto vacío*, *conjunto*. 29

conjunto vacío El conjunto vacío \emptyset es el conjunto que carece de elementos. Por lo tanto, su cardinalidad es 0. Ver *cardinalidad*, *conjunto*. 27

grado Número de nodos a los cuales se conecta un nodo particular en un grafo. Si un nodo A se conecta a otros seis, el grado k de A se expresa como $k_A = 6$. Ver *grafo*. 2, 3, 10, 12, 13, 27

grafo Objeto matemático que permite ilustrar y modelar sistemas complejos según las relaciones entre sus componentes, representados por símbolos llamados *nodos*. Estas relaciones se representan a su vez como líneas llamadas *aristas* que conectan únicamente pares de nodos. Formalmente, un grafo es un par ordenado $G = (N, A)$, donde N es el conjunto de los nodos y A es un conjunto de pares ordenados de elementos de N , que representa el conjunto de aristas. Ver *par ordenado, conjunto*. 9, 11, 12, 14, 16, 17, A-1

homeostasis Equilibrio entre un organismo y su entorno que permite la supervivencia del organismo. Contrasta con el equilibrio termodinámico en general en que no puede ser estático; el organismo necesita nutrientes y condiciones externas (temperatura, pH) e internas (equilibrio hídrico y de otras concentraciones, estabilidad de las macromoléculas) para evitar la muerte el tiempo suficiente para reproducirse. 1, 15, 17

jerárquica Topología en la cual de nodos antecesores surgen conexiones a nodos subordinados, y de éstos a otros subordinados. Se distingue de los árboles porque las conexiones no son sólo antecesor–subordinado; los subordinados pueden conectarse entre sí, pero las aristas clave, ya sea por número o dirección, dan a los antecesores. Ver *topología*. 10

NP-completo Tipo de problema matemático, clasificado según la dificultad de cálculo que implica resolverlo. Los problemas NP-completos son, al momento de escribir este trabajo, lentos de resolver, porque aunque verificar una solución es rápido, calcular las soluciones mismas es lento. 16

NP-duro Tipo de problema matemático, clasificado según la dificultad de cálculo que implica resolverlo. Los problemas NP-duros no son, al momento de escribir este trabajo, rápidos o incluso posibles de resolver en todos los casos [36]. 16

par ordenado Conjunto de dos elementos en el cual el orden tiene un significado. El ejemplo estándar es un punto en un plano euclidiano con ejes, en el cual cada elemento designa posiciones respecto de los ejes y el ordenamiento designa a qué eje corresponde cada elemento. Formalmente, un par ordenado (u, v) es un conjunto definible como $(u, v) = \{\{u\}, \{u, v\}\}$. Ver *conjunto*. 29

producto Cartesiano Conjunto basado en dos conjuntos A y B y denotado por $A \times B$, cuyos elementos son todos los pares ordenados (a, b) tales que $a \in A$ y $b \in B$. Ver *par ordenado, conjunto*. 27–29

red Ver *grafo*. 2, 9–13, 18, 25–27

ruta metabólica Conjunto de reacciones, comúnmente catalizadas por enzimas, que median secuencialmente la transformación de un compuesto a otro en el metabolismo. Las rutas estándar se representan como caminos entre compuestos de importancia; un ejemplo son las rutas de síntesis de aminoácidos. 1, 3, 6, 7, 13, 15–19, 47, 57

topología Acomodo de todos los nodos en un grafo. La topología de un grafo puede verse en la representación gráfica de éste, lo cual genera formas diferentes que son útiles para diferenciar cada topología de las demás. Ver *grafo*. 12, 13, 27, 36, 39, 54, 55

Siglas

NH₄ Amonio. 1–3, 17, 23

PO₄ Ortofosfato. 1, 2, 23, 30

ADP Adenosina–5'–difosfato. 1–3, 7, 8, 23, 25, 30

ATP Adenosina–5'–trifosfato. 1–3, 7, 8, 12, 17, 22, 23, 25, 26, 30

CoA Coenzima A. 1, 2, 22, 23

CTS *cluster of tree structures*. 36, 39, 41–45, 49–58, A-12

EC *Enzyme Comission Number*. 4–7, 44, 45, 49–56, 58, A-12

ECP Elemento del producto Cartesiano. 27–33, 39

FAD Flavina–adenina dinucleótido. 22

FBA *flux balance analysis*. 15, 16

FMN Flavina mononucleótido. 7

Gln Glutamina. 17, 23

Glu Ácido glutámico. 3, 17, 23

KEGG *Kyoto Encyclopedia of Genes and Genomes*. 5–8, 14, 17, 28, 42, 47, 49, 58

NAD Nicotina–adenina dinucleótido. 1, 2, 22, 23, 30, 36, 52, 53

NADH Nicotina–adenina dinucleótido reducido. 1, 2, 23, 30

NADP NAD–fosfato. 1, 2, 36, 52, 53

NADPH NAD–fosfato reducido. 1, 2

TCA ciclo de los ácidos tricarboxílicos. 1, 3, 17

TS *tree structure*. 33, 34, 36, 37, 39, 41, 42, 45, 47, 58

Bibliografía

- [1] M Arita. The metabolic world of *Escherichia coli* is not small. *Molecular Biosystems*, 5:1482–1493, 2009.
- [2] M Arita. *From Metabolic Reactions to Networks and Pathways*, chapter 6. Springer Science + Business Media (US), 2012.
- [3] AL Barabási and ZN Oltvai. Network biology: Understanding the cell’s functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.
- [4] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57(1):289–300, 1995.
- [5] JM Berg, JL Tymoczko, and L Stryer. *Biochemistry*. W H Freeman, 5th edition, 2002. ISBN 9781429276351 1429276355.
- [6] NH Bergman, editor. *Gene Annotation and Pathway Mapping in KEGG*, volume 2, chapter 6. Humana Press Inc., 2007.
- [7] A Chowdhury and CD Maranas. Designing overall stoichiometric conversions and intervening metabolic reactions. *Scientific Reports*, 5(16009), 2015.
- [8] DA Cuevas, J Edirisinghe, CS Henry, R Overbeek, TG O’Connell, and RA Edwards. From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model. *Frontiers in Microbiology*, 7:907, 2016.
- [9] K Faust, D Croes, and J van Helden. Metabolic Pathfinding Using RPAIR Annotation. *Journal of Molecular Biology*, 388:390–414, 2009.
- [10] K Faust and J van Helden. *Predicting Metabolic Pathways by Sub-Network Extraction*, chapter 7, pages 107–130. 2012.

- [11] CM Fennesey, SE Ivie, and MS McClain. Coenzyme depletion by members of the aerolysin family of pore-forming toxins leads to diminished ATP levels and cell death. *Molecular bioSystems*, 8(8):2097–2105, 2012.
- [12] AJ Gates and LM Rocha. Control of complex networks requires both structure and dynamics. *Scientific Reports*, 6(24456), 2016. doi: 10.1038/srep24456.
- [13] P Gerlee, L Lizana, and K Sneppen. Pathway identification by network pruning in the metabolic network of *Escherichia coli*. *Bioinformatics*, 25(24):3282–3288, 2009.
- [14] M Hattori, Y Okuno, S Goto, and M Kanehisa. Heuristics for Chemical Compound Matching. *Genome Informatics*, 14:144–153, 2003.
- [15] M Hattori, N Tanaka, M Kanehisa, and S Goto. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Research*, 38(suppl 2):W652–W656, 2010.
- [16] AP Heath, GN Bennett, and LE Kavraki. Finding metabolic pathways using atom tracking. *Sample Journal*, 26(12):1548–1555, 2010.
- [17] M Heinonen, S Lappalainen, T Mielikäinen, and J Rousu. Computing atom mappings for biochemical reactions without graph isomorphism. *Journal of Computational Biology*, 18(1):43–58, 2011.
- [18] L Hertz and TM Jeitner. Glutamine-Glutamate Cycle Flux Is Similar in Cultured Astrocytes and Brain and Both Glutamate Production and Oxidation Are Mainly Catalyzed by Aspartate Aminotransferase. *Biology (Basel)*, 6(17), 2017. doi:10.3390/biology6010017.
- [19] Y Huang, C Zhong, HX Lin, and J Wang. A Method for Finding Metabolic Pathways Using Atomic Group Tracking. *PLOS ONE*, 12(1), 2017.
- [20] H Jeong, B Tombor, R Albert, ZN Oltvai, and AL Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.
- [21] M Kanehisa, S Goto, S Kawashima, Y Okuno, and M Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32:D277–D280, 2004.
- [22] M Kanehisa, Y Sato, M Kawashima, M Furumichi, and M Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44:D457–D462, 2016. doi: 10.1093/nar/gkv1070.

- [23] KJ Kauffman, P Prakash, and JS Edwards. Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(1):491–496, 2003.
- [24] O Khersonsky and DS Tawfik. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry*, 79:471–505, 2010.
- [25] HS Kim, JE Mienthal, and G Caetano-Anollés. MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics*, 7(351), 2006.
- [26] M Kotera, Y Okuno, M Hattori, S Goto, and M Kanehisa. Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions. *Journal of the American Chemical Society*, 126(50):16487–16498, 2004. doi: 10.1021/ja0466457.
- [27] A Kun, B Papp, and E Szathmáry. Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biology*, 9(R51), 2008.
- [28] M Lakshmanan, G Koh, BKS Chung, and DY Lee. Software applications for flux balance analysis. *Briefings in Bioinformatics*, 15(1):108–122, 2012.
- [29] M Latendresse, JP Malerich, M Travers, and PD Karp. Accurate atom-mapping computation for biochemical reactions. *Journal of Chemical Information and Modeling*, 52:2970–2982, 2012.
- [30] DARS Latino and J Aires-de Sousa. *Classification of Chemical Reactions and Chemoinformatic Processing of Enzymatic Transformations*, pages 325–340. Humana Press, Totowa, NJ, 2011.
- [31] G Lima-Méndez and J van Helden. The powerful law of the power law and other myths in network biology. *Molecular Biosystems*, 5:1482–1493, 2009.
- [32] H Lu, B Shi, G Wu, Y Zhang, X Zhu, Z Zhang, C Liu, Y Zhao, T Wu, J Wang, and R Chen. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochemical and Biophysical Research Communications*, 345:302–309, 2006.
- [33] X Mao, T Cai, JG Olyarchuk, and L Wei. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 21(19):3787–3793, 2005.

- [34] Y Moriya, M Itoh, S Okuda, AC Yoshizawa, and M Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35:W182–W185, 2007.
- [35] A Muto, M Kotera, T Tokimatsu, Z Nakagawa, S Goto, and M Kanehisa. Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. *Journal of Chemical Information and Modeling*, 53(1):613–622, 2013.
- [36] Z Nikoloski, S Grimbs, P May, and J Selbig. Metabolic networks are NP-hard to reconstruct. *Journal of Theoretical Biology*, 254:807–816, 2008.
- [37] E Pitkänen, P Jouhten, and J Rousu. Inferring branching pathways in genome-scale metabolic networks. *BMC Systems Biology*, 3(103), 2009.
- [38] FJ Planes and JE Beasley. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Briefings in Bioinformatics*, 9(5):422–436, 2008.
- [39] SA Rahman, S Martinez-Cuesta, N Furnham, GL Holliday, and JM Thornton. EC-BLAST: A Tool to Automatically Search and Compare Enzyme Reactions. *Nat Methods*, 11(2):171–174, 2014.
- [40] E Ravasz, AL Somera, DA Mongru, ZN Oltvai, and AL Barabasi. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297:1551–1555, 2002.
- [41] G Rodrigo, J Carrera, K Jones Prather, and A Jaramillo. DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics*, 24(21):2554–2556, 2008.
- [42] Y Shimizu, M Hattori, S Goto, and M Kanehisa. Generalized reaction patterns for prediction of unknown enzymatic reactions. *International Conference on Genome Informatics*, 20:149–158, 2008.
- [43] MD Stobbe, GA Jansen, PD Moerland, and AHC van Kampen. Knowledge representation in metabolic pathway databases. *Briefings in Bioinformatics*, 15(3):455–470, 2014.
- [44] K Tipton and S Boyce. History of the enzyme nomenclature system. *Bioinformatics*, 16:34–40, 2000.

- [45] A Varma and BO Palsson. Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic By-Product Secretion in Wild-Type *Escherichia coli* W3110. *Applied and Environmental Microbiology*, 60(10):3724–3731, 1994.
- [46] AR Zomorodi and D Segrè. Synthetic Ecology of Microbes: Mathematical Models and Applications. *Journal of Molecular Biology*, 428:837–861, 2016.