VNIVERSIDAD NACIONAL AVTNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"Alineamiento múltiple de vías metabólicas usando cómputo evolutivo"

T E S I S

Que para obtener el grado de
Maestra en Ciencias de la Computación

PRESENTA:

PATRICIA GUADALUPE ORTEGÓN CANO

Directora de Tesis: Dra. Katya Rodríguez Vázquez
Co-Director de Tesis: Dr. Ernesto Pérez Rueda

MÉXICO, DF, 2011

Contenido

Resumen 6
Abstract 7
Capítulo 1 INTRODUCCIÓN 8
1.1 Bioinformática y cómputo evolutivo 8
1.2 Problema y Justificación 9
1.3 Objetivos 10
1.4 Estructura del trabajo 11
Capítulo 2 INTRODUCCIÓN AL METABOLISMO 12
2.1 Definiciones 12
2.2 Números enzimáticos 13
2.3 Evolución 14
2.3.1 Evolución de las vías metabólicas 15
2.4 Análisis de las vías metabólicas 18
Capítulo 3 ANTECEDENTES 20
3.1 Análisis vías metabólicas 20
3.2 Alineamiento de secuencias biológicas 20
3.2.1 Cómputo Evolutivo y Alineamiento de Secuencias 23
3.3 Alineamiento de vías metabólicas 25
3.3.1 Algoritmos basados en la topología de las vías 25
3.3.2 Algoritmos basados en la composición de las vías 27
Capítulo 4 METODOLOGÍA 34
4.1 Base de datos 34
4.2 Representación de las vías metabólicas 36
4.2.1 *Breadth First Search* 37
4.3 Algoritmo Genético 37
4.3.1 Representación 39
4.3.2 Operadores 40
4.3.2.1 Selección 40
4.3.2.2 Cruza 41
4.3.2.3 Mutación 43
4.3.3 Evaluación de los alineamientos 44
4.3.4 Función objetivo 45
4.3.4.1 Entropía mínima 50
4.3.4.2 Suma de pares 52
4.4 Algoritmo progresivo 53
4.5 Análisis evolutivo de las secuencias 54
4.5.1 *Clustering* 54
4.5.1.1 *K-Means* 54
4.6 Visualizador 56
Capítulo 5 RESULTADOS 57

5.1 Vías metabólicas y secuencias 57
5.2 Alineamientos por pares 58
5.3 Clustering 61
5.4 Alineamientos múltiples 66
5.5 Mapeo de los alineamientos a las vías metabólicas. 71
Capítulo 6 CONCLUSIONES 73
Capítulo 7 PERSPECTIVAS 75
Apéndice A 76
Apéndice B 78
Apéndice C 89
Apéndice D 90
Apéndice E 91
Bibliografía 104

Agradecimientos

Más vale tarde que nunca...me tardé un poco mas de lo planeado pero finalmente aquí está la tan esperada tesis de maestría...

Primero que nada quisiera agradecer a mi familia por su apoyo en todos mis planes y proyectos.

También a mis amigos, el inolvidable “bombers team”, Fátima, Rodri, Ireri, por todos los momentos buenos, malos, divertidos y estresantes que pasamos en esta aventura de poco mas de dos años.

A la Dra. Katya Rodríguez Vázquez y al Dr. Ernesto Pérez Rueda por su confianza, apoyo, comentarios y ayuda para lograr este trabajo.

Y a todas las personas que tuve la oportunidad de conocer durante la realización de esta tesis, profesores y amigos, que con sus comentarios y ayuda contribuyeron de alguna manera a este trabajo.

Resumen

La comparación de secuencias es una de las tareas más comunes en la bioinformática. Sin embargo, la mayoría de las herramientas existentes son para secuencias de aminoácidos y ADN. Las vías metabólicas son consideradas las unidades funcionales de los sistemas biológicos y el alineamiento de estas vías es una de las herramientas más poderosas para el análisis comparativo del metabolismo.

La complejidad de la información involucrada en el análisis de las vías metabólicas implica la necesidad de métodos computacionales eficientes.

La computación evolutiva y específicamente, los algoritmos genéticos (AGs) han sido aplicados con éxito en diversos problemas de la ingeniería. Ofrecen una clara separación entre el criterio de evaluación (función objetivo) y el proceso de optimización, y además son altamente paralelizables. Estas son algunas de las razones por las cuales últimamente han sido aplicados a problemas de la bioinformática.

En esta tesis, un enfoque del cómputo evolutivo es utilizado para el alineamiento de vías metabólicas. En la representación propuesta, las vías son transformadas a secuencias de enzimas y estas secuencias son alineadas por el AG. Se presenta un criterio para evaluar la calidad del alineamiento, basado en los conceptos de la teoría de la información.

El AG implementado fue altamente eficiente al ser combinado con la técnica del alineamiento progresivo. El enfoque propuesto nos permite visualizar fácil y eficientemente las similitudes y diferencias entre regiones completas de las vías metabólicas, lo cual es novedoso en el área de la comparación y evolución del metabolismo.

Se realizaron experimentos con los alineamientos de secuencias de 47 vías metabólicas pertenecientes a *E. coli*, identificando regiones de vías metabólicas que comparten una sucesión de pasos metabólicos similares, lo cual sugiere una catálisis común.

Estos resultados permiten hacer inferencias a cerca del proceso evolutivo dentro de las vías metabólicas.

Abstract

Sequence comparison is one of the most common tasks in bioinformatics. Nevertheless, the existent tools are mainly for amino acids and DNA sequences. Metabolic pathways are considered as the functional units of biological systems, and the alignment of these pathways is one of the most powerful tools for comparative analysis of metabolism.

The complexity of the information involved in the analysis of metabolic pathways implies the necessity for efficient computational methods.

Evolutionary computation and specifically, genetic algorithms (GAs) have been successfully applied to a variety of engineering problems. They offer a clearly separation between the evaluation criteria (objective function) and the optimization process, and also are easy to parallelize**.** These are some of the reasons why lately GAs have been applied in bioinformatics.

In this thesis, an evolutionary computing approach is used for the alignment of metabolic pathways. In the proposed representation, pathways are transformed to sequences of enzymes and these sequences are aligned by the GA. A criterion to assess the quality of the sequence alignment based on information theory concepts is also introduced.

The implemented GA was highly efficient combined with the progressive alignment technique. The proposed approach allow us easily and efficiently visualize the similarities and the differences between completed regions of metabolic pathways, which is a novelty in the analysis of comparison and evolution of metabolism.

Experiments were conduced with alignments of sequences from 47 pathways belonging to *E. coli*, identifying regions of metabolic pathways sharing a similar succession of metabolic steps, suggesting common catalysis, which are non-trivial to identify with traditional computational tools.

These results allow us to make inferences about the evolutionary process in the metabolic pathways.

Capítulo 1

INTRODUCCIÓN

1.1 Bioinformática y cómputo evolutivo

La bioinformática puede ser vista como el *uso de métodos computacionales para hacer descubrimientos biológicos*. Es un campo interdisciplinario que involucra a la biología, las ciencias de la computación, las matemáticas y la estadística, y que tiene por objetivo el análisis exhaustivo y sistematizado de grandes cantidades de información, tales como secuencias biológicas, contenido y distribución de genomas, así como la predicción de la función y estructura de macromoléculas. El objetivo final del campo es permitir la generación de nuevos principios biológicos, y de proveer los medios para obtener respuestas a preguntas de inmensa relevancia para las ciencias de la vida.

Recientemente, los algoritmos evolutivos (AEs), una clase de técnicas de búsqueda aleatoria y optimización guiadas por los principios de la evolución y genética, han ganado la atención de investigadores para la solución de problemas bioinformáticos. Los algoritmos genéticos (AGs), las estrategias evolutivas (EE), y la programación genética (PG) son los principales representantes de los AEs. De estas técnicas, los AGs son procesos de búsqueda eficientes, adaptativos y robustos, que producen soluciones muy cercanas a las soluciones óptimas, y son altamente paralelizables, por lo que son usados ampliamente.

Anteriormente, las herramientas de análisis de datos usadas para el análisis de secuencias estaban basadas en técnicas estadísticas como la regresión y la estimación. El papel de los AGs en la bioinformática ha ganado importancia dada la necesidad de manejar grandes conjuntos de datos en biología en una forma robusta y computacionalmente eficiente.

La velocidad a la que la información biológica se ha incrementado en los últimos años demanda el uso de técnicas más eficientes para analizarla y organizarla. Las técnicas de cómputo evolutivo han sido probadas en diversos campos de la ingeniería con resultados

muy eficientes trabajando con grandes espacios de búsqueda. Gracias a todas las ventajas que ofrecen, como son la facilidad para paralelizar los procesos, o la adaptación de la función a optimizar, últimamente han sido aplicadas a problemas en el campo de la bioinformática, en el reconocimiento de patrones en grandes bases de datos, en problemas de optimización de funciones de energía, para la predicción de estructuras conformacionales y para el alineamiento y comparación de secuencias biológicas [43].

La secuenciación de genomas de varios organismos ha traído a la comunidad científica una vasta cantidad de datos a partir de los cuales es posible obtener información acerca de las secuencias y estructuras de varios miles de proteínas y genes. Adicionalmente, esta información ha permitido generar nuevo conocimiento sobre la interacción de estas mismas entidades, como son las redes regulatorias de genes, redes de interacción de proteínas y redes metabólicas, entre otras [14,23].

Las aplicaciones computacionales en el campo biológico están ganando una gran importancia para el análisis de la información antes mencionada, ya que el conocimiento obtenido por estos análisis es necesario en el diseño de fármacos, en la obtención de vacunas y en el estudio de nuevas enfermedades. Este análisis siempre va acompañado de experimentación en laboratorios, pero en los últimos años se está apoyando más en el estudio e investigación *in silico* (usando simulaciones computacionales), para abaratar el costo en investigación. Es decir, con este tipo de enfoques se pueden diseñar experimentos puntuales para corroborar una predicción en particular [8].

1.2 Problema y Justificación

La generación de esta cantidad de información tiene como consecuencia la creación de enormes bases de datos disponibles a la comunidad científica, tales como las bases de datos metabólicas: KEGG, ECOCYC, BIOCYC, y METACYC [2]. Aunque el estudio del metabolismo no es reciente, la creación de estas bases de datos y las posibilidades que ofrecen sí lo es.

Aún cuando la información proporcionada por los genomas secuenciados puede dar indicios sobre su evolución y metabolismo celular, el conocimiento aislado del genoma solo es el punto de inicio del trabajo real. En este sentido, las unidades funcionales de los sistemas biológicos son las vías metabólicas. De hecho, los esfuerzos de investigación comienzan a fijar su atención hacia esta área conforme la disposición de la información antes mencionada hace posible la comparación de las vías metabólicas [14].

La comparación de secuencias es una de las tareas más comunes en la bioinformática, sin embargo la mayoría de las herramientas existentes para realizar esta tarea son para secuencias de proteínas y ADN. El *alineamiento de vías metabólicas* es una de las herramientas más poderosas para el análisis comparativo del metabolismo. Observando el mapa de todas las vías metabólicas conocidas (Figura 2.3), es posible darse cuenta de la complejidad que implica realizar el análisis de toda esta información, y de la necesidad implícita de métodos computacionales eficientes. Estas herramientas permiten corroborar los resultados encontrados experimentalmente y también hacer predicciones.

Las propuestas existentes para la solución de este problema son relativamente pocas y diversas. En general, cada una tiene un enfoque diferente en el algoritmo de alineamiento, así como en el método para la reconstrucción de las vías a partir de las bases de datos. La mayoría de estos trabajos se enfoca principalmente en el alineamiento por pares de las vías metabólicas.

1.3 Objetivos

En este trabajo, el problema de alineamiento de vías metabólicas es considerado desde la perspectiva de las enzimas con el objetivo de encontrar relaciones evolutivas entre diferentes vías y metabolismos.

Dentro de los objetivos específicos está la propuesta de una representación de las vías metabólicas diferente a las existentes, que nos permitirá identificar visualmente la información similar entre las secuencias, lo cual es muy difícil de lograr con algunos de los métodos propuestos actualmente. Para realizar los alineamientos múltiples se desarrollará un algoritmo genético que, gracias a la flexibilidad que nos ofrecen estos algoritmos,

permite utilizar diversas funciones para calificar la calidad de los alineamientos. Se propone una función que utiliza conceptos de entropía y teoría de la información para evaluar los alineamientos obtenidos con el algoritmo genético.
Para lograr una mejor calidad en los alineamientos múltiples se propone un método para la creación de grupos de secuencias.

1.4 Estructura del trabajo

El capítulo 2 introduce algunos de los conceptos básicos del metabolismo involucrados en el problema que se está presentando. En el capítulo 3 se explica a más detalle este problema junto con una revisión del estado del arte en el área. En el capítulo 4 se describe la metodología que se propone para la solución de este problema; en los capítulos 5 y 6, se muestran los resultados obtenidos y se presenta una breve discusión y conclusiones. Finalmente, en el capítulo 7 se presentan las perspectivas para un trabajo futuro.

Capítulo 2

INTRODUCCIÓN AL METABOLISMO

2.1 Definiciones

A nivel celular ocurren miles de reacciones químicas, el conjunto de todas estas reacciones es conocido como metabolismo, el cual se divide en dos partes:

Catabolismo: Conjunto de reacciones degradativas de los nutrientes, por las que se obtiene energía, poder reductor y moléculas precursoras de las macromoléculas biológicas.

Anabolismo: Conjunto de reacciones biosintéticas que requieren energía y poder reductor.

Todas estas reacciones químicas se encuentran interconectadas en una gigantesca red finamente regulada. Los compuestos que forman parte de las reacciones o que son formados a partir de ellas son llamados **metabolitos**. Las "entradas" son conocidas como **sustratos** y las "salidas" son los **productos**. Para que estas reacciones ocurran en la célula se debe superar cierto valor de energía de activación. Los catalizadores son los encargados de hacer que este valor sea alcanzado, creando las condiciones necesarias en el medio.

Las **enzimas** son proteínas que funcionan como catalizadores de eficiencia y especificidad primaria, es decir que solo actúan sobre determinados sustratos y producen solo un tipo de reacción. La importancia de las enzimas es tal, que muchas de las reacciones que ellas catalizan no se pueden llevar a cabo si la enzima está ausente [32].

En la Figura 2.1 podemos observar la representación general de una reacción química, con sus tres componentes principales, el sustrato (S), la enzima (E), y finalmente después de la transformación se obtiene el producto (P) final.

$$E + S \rightleftharpoons [ES] \rightleftharpoons E + P$$

Figura 2.1. Representación de una reacción química.

Estas reacciones se pueden agrupar en unidades funcionales denominadas **vías o rutas metabólicas**. El conjunto de todas las vías metabólicas de un organismo es conocido como

una **red metabólica**. Las vías metabólicas pueden verse como secuencias de reacciones que se efectúan una tras otra de tal manera que el producto de una reacción catalizada por una enzima específica, es el sustrato de la siguiente reacción. Entonces, tenemos para cada reacción una enzima asociada, lo cual nos permite realizar simplificaciones en la representación de las vías para su análisis [14].

Así, tenemos vías metabólicas para degradar los lípidos, para degradar azúcares, para sintetizar nucleótidos, entre otras. En la Figura 2.2 se muestra un segmento de la vía de la glucólisis [32].

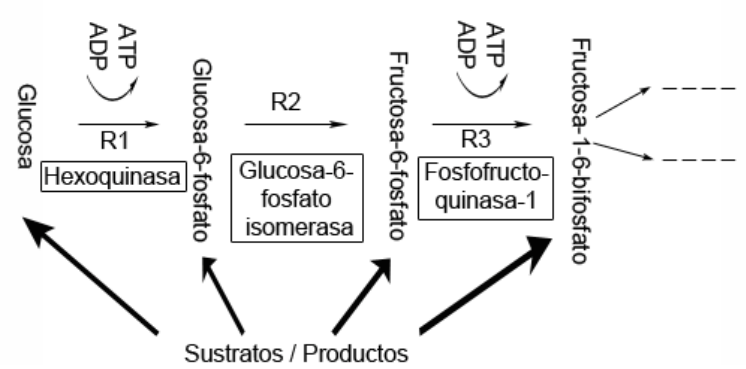


Figura 2.2. Segmento de la vía de la glucólisis. Esta vía metabólica transforma la glucosa en energía. Podemos observar que hay tres reacciones en el segmento (R1, R2 y R3), y cada una tiene una enzima asociada (rectángulos), el producto de R1 es el sustrato de R2 y así sucesivamente.

2.2 Números enzimáticos

Anteriormente las enzimas eran nombradas atendiendo al sustrato sobre el cual actuaban, añadiendo el sufijo que hacía referencia a la reacción catalizada. Debido al gran número de enzimas conocidas en la actualidad, se ha adoptado una clasificación y nomenclatura más sistemática, en la que cada enzima tiene un número de clasificación que la identifica. Estos números son asignados por el Comité de la Unión Internacional de Bioquímica. Están formados por cuatro dígitos separados por puntos (A.B.C.D), cada nivel es una clasificación de la enzima, que se realiza de acuerdo a la reacción química que cataliza.

A: Clase, tipo de reacción

B: Subclase, indica el sustrato

C: Sub sub clase, indica el cosustrato

D: Número de orden.

Así, el primer nivel clasifica a las enzimas en seis grupos de acuerdo a la clase general de reacciones que catalizan:

1 **Óxido-reductasas.** Catalizan las reacciones de oxidación-reducción.
2 **Transferasas.** Catalizan reacciones de transferencia de grupos.
3 **Hidrolasas.** Catalizan la hidrólisis.
4 **Liasas.** Catalizan las reacciones de eliminación no hidrolítica, no oxidante de un sustrato en reacciones que generan un enlace doble.
5 **Isomerasas.** Catalizan reacciones de isomerización.
6 **Ligasas.** Catalizan la ligadura o unión de dos sustratos en reacciones sintéticas que requieren de energía química potencial.

De esta forma la enzima representada por el número enzimático **3.1.6.5** tiene las siguientes características:
Nivel 1 nos dice que es una hidrolasa, es decir la reacción que realiza es la hidrólisis.
Nivel 2 actúa sobre los enlaces éster.
Nivel 3 su aceptor es el éster sulfúrico.
Nivel 4 elimina una entrada.

El hecho de que dos enzimas tengan en los subniveles (2, 3 y 4) el mismo número, no significa que actúen sobre el mismo sustrato o que tengan el mismo aceptor para la reacción. Así, la enzima 3.1.6.5 es una hidrolasa que actúa sobre los enlaces éster y la enzima 1.1.5.1 es una enzima óxido-reductasa que actúa sobre el grupo de donadores OH-CH, es decir, a pesar de que ambas tienen el segundo nivel igual a 1 actúan sobre distintos componentes. Los primeros dos niveles son considerados los más importantes para determinar si dos enzimas catalizan reacciones químicamente similares [1].

2.3 Evolución

Uno de los enfoques principales del análisis de la información biológica es el evolutivo, debido a que nos permite entender la presencia o ausencia de ciertos elementos en

organismos que en primera instancia no se encuentran relacionados. Una de las principales herramientas para estos análisis es la comparación de secuencias de nucleótidos y de aminoácidos, usando alineamientos. El objetivo es reconocer segmentos que se han conservado a lo largo de la evolución en un conjunto de secuencias. Esta identificación se realiza alineando estas secuencias, permitiendo localizar secciones repetidas así como las diferencias existentes.

En 1973, el biólogo evolutivo Theodosius Dobzhansky señaló que "nada en la biología tiene sentido sin la luz de la evolución". Los biólogos han reconocido que la evolución es la clave para el entendimiento del desarrollo de la vida en la Tierra. De hecho, la evolución ha sido vista por los biólogos, como un proceso de aprendizaje desde principios de 1930. Actualmente estamos en la era donde nuestra capacidad para aprovechar la evolución como una herramienta de la ingeniería en el laboratorio es tan interesante como el significado de la evolución en los sistemas naturales [22].

2.3.1 Evolución de las vías metabólicas

Todos los organismos poseen una red que conecta todas las rutas metabólicas, las relacionadas con la biosíntesis de los bloques constructores de proteínas, ácidos nucleícos, lípidos y carbohidratos, y las involucradas en el catabolismo de diferentes compuestos que manejan los procesos celulares (Figura 2.3).

El cómo estas rutas se originaron y evolucionaron se ha discutido por décadas y el debate aún continúa. Se han desarrollado varias teorías para explicar la evolución de las enzimas de una vía metabólica a partir de los constituyentes de una sopa prebiótica, las ideas principales están basadas en la duplicación de genes [30].

En este sentido, el modelo de la *evolución retrógrada* propuesto por Horowitz en 1945 y el *modelo evolutivo patchwork* de Jensen en 1976 han sido los modelos más aceptados para explicar la evolución del metabolismo. En el primero, las enzimas evolucionan "hacia atrás" con respecto a la dirección de la vía, cuando un sustrato se agota en el medio, la duplicación génica puede generar una enzima capaz de proveer ese sustrato, dando lugar a reacciones consecutivas catalizadas por enzimas codificadas por esos duplicados [47].

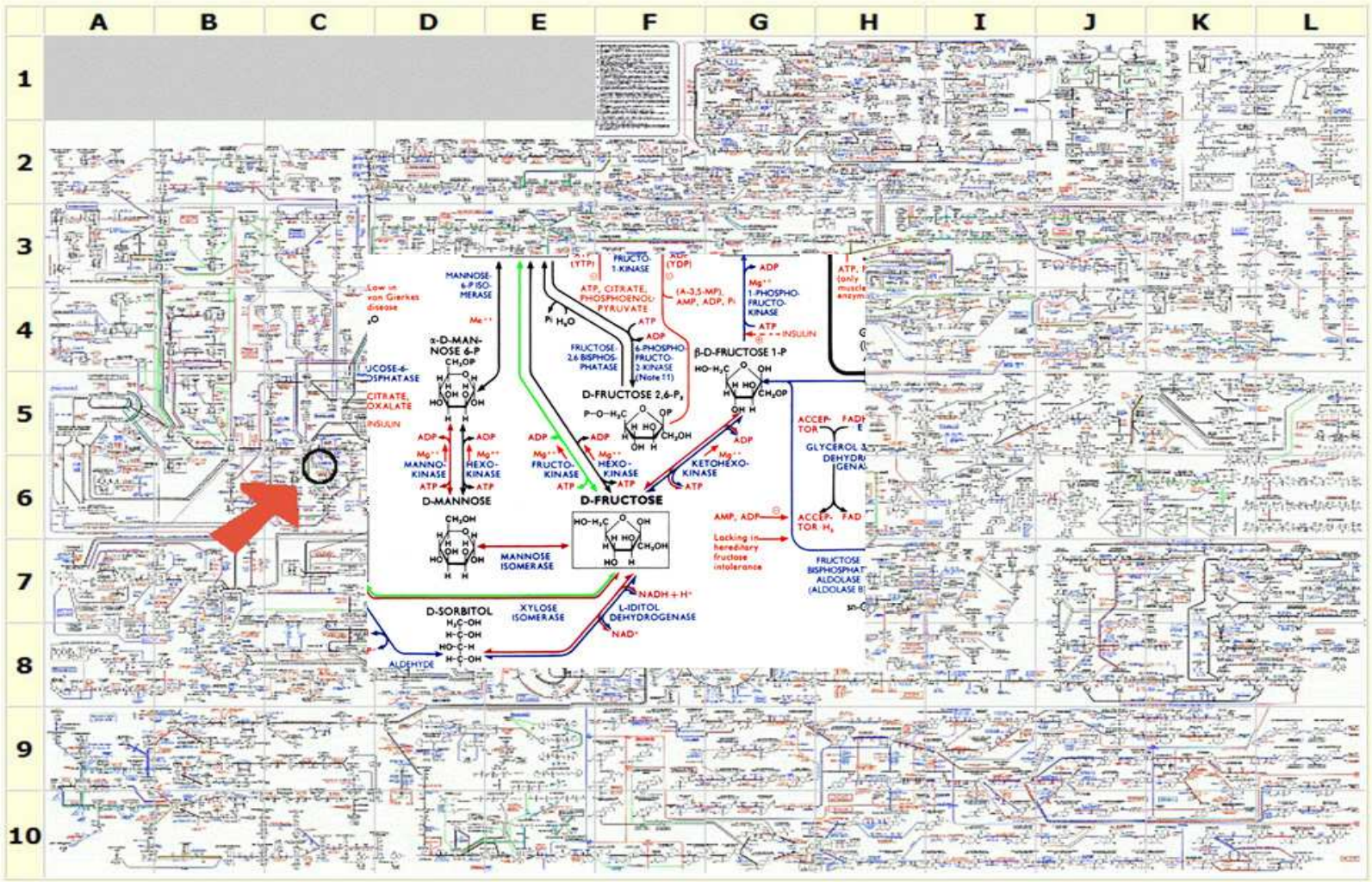


Figura 2.3. Mapa del conjunto de vías metabólicas conocidas. Al hacer un acercamiento sobre la imagen se puede observar con más detalle cada una de las reacciones que participan junto con los compuestos[1].

[1] http://www.expasy.ch/cgi-bin/show_thumbnails.pl

Para ilustrar este modelo consideremos el siguiente ejemplo: la enzima **E1** cataliza la reacción **A→B**, en la cual **B** es esencial para el organismo; ahora supongamos que **A** empieza a agotarse en el ambiente, esto significa que un organismo que hospede a una enzima **E2** que pueda catalizar una reacción que produzca **A** de cualquier otro sustrato tendrá una ventaja considerable. Puesto que **E1** acepta a **A** como sustrato, tiene una gran oportunidad de sobresalir entre las enzimas que no tienen afinidad por **A** y por lo tanto podrá ser duplicada y mutada en **E2**. **E2** será rápidamente aceptada y cualquier mutación nula que involucre a **E2** será letal, por lo que se favorece la preservación de esta enzima en la vía [47]. En la Figura 2.4 se explica este ejemplo.

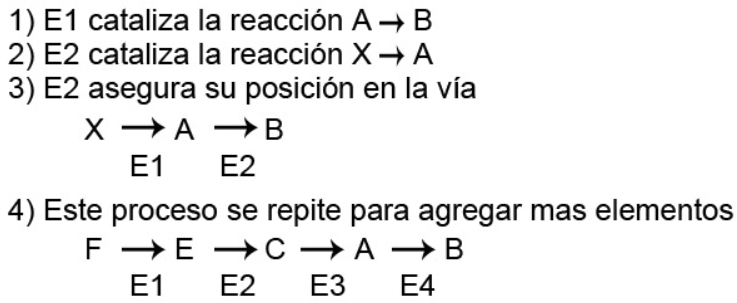


Figura 2.4. Ejemplo de la evolución de la vía metabólica usando el modelo de la evolución retrógrada [36]**.**

En el modelo evolutivo por *patchwork*, también conocido como "evolución por reclutamiento", las enzimas presentan una amplia especificidad de sustratos, a partir de ahí fueron evolucionando y la duplicación de genes explica la posterior especificación. Supongamos que la enzima **E** cataliza una reacción en la que se aceptan los sustratos **S1** y **S2**. A través de la duplicación y la mutación aleatoria evolucionan dos versiones diferentes de la enzima **E**, **E'** que acepta el sustrato **S1** y **E"** que únicamente acepta a **S2** como sustrato [35,47]. Esto se ilustra en la Figura 2.5.

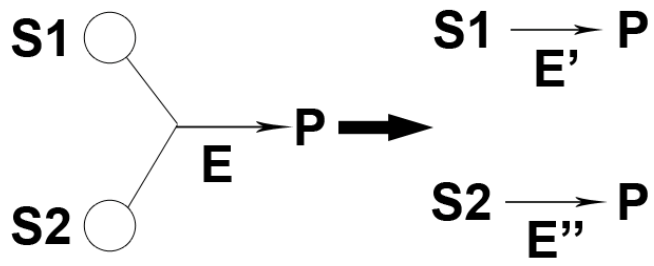


Figura 2.5. Ejemplo de la evolución de la vía metabólica usando el modelo de evolución *patchwork* [36]**.**

Dado que el modelo de evolución *retrógrada* involucra reacciones consecutivas, se considera que puede generar reacciones químicamente diferentes, preservando las propiedades de unión por el tipo de sustrato. En contraste, se ha sugerido que el modelo *patchwork* puede generar enzimas que catalizan reacciones químicamente similares, aún cuando actúen sobre sustratos de diferentes tipos. Una manera sencilla de determinar si dos enzimas catalizan reacciones químicamente similares o no, es comparando los primeros dos dígitos de sus números enzimáticos (EC:a.b.-.-). Varios autores han usado las diferencias entre estos modelos para tratar de determinar su contribución en la evolución del metabolismo, señalando al modelo *pathwork* como el predominante [18].

2.4 Análisis de las vías metabólicas

La comparación de las vías metabólicas tiene como objetivo, identificar similitudes entre ellas, y entre los metabolismos de diversos organismos, lo cual proporciona ideas para la identificación de vías alternas, reconstrucción de árboles filogenéticos, entre otras. El análisis de estas vías también es de relevancia para identificar los blancos de nuevos fármacos así como de otras aplicaciones biotecnológicas [16].

Durante la evolución de las vías metabólicas pueden ocurrir los siguientes eventos:

- **Deleción**. Se pierden algunas enzimas debido a que se dejan de usar.
- **Mutación**. Algunas enzimas son sustituidas por otras que catalizan una reacción similar (*patchwork*).
- **Inserción**. Se introducen nuevas enzimas para compensar las reacciones no catalizadas o el agotamiento de sustratos en el medio ambiente (retrógrada).

Al analizar las vías metabólicas podemos descubrir estos eventos que ocurren a través del proceso evolutivo entre grupos de vías. La forma más sencilla de encontrar estas diferencias es mediante un análisis comparativo de vías metabólicas alineando unas con otras.

Como se mencionó anteriormente, al centrar el análisis en las enzimas, podemos simplificar la representación de las vías como una secuencia lineal de estas enzimas, usando los números enzimáticos para representarlas. De esta forma, aplicando ciertas transformaciones, podemos obtener secuencias de números enzimáticos que son la representación simplificada de las vías metabólicas, las cuales permiten aplicar el enfoque clásico de alineamiento de secuencias para la búsqueda de relaciones evolutivas entre ellas.

Capítulo 3

ANTECEDENTES

3.1 Análisis vías metabólicas

Existen diversos enfoques para el análisis de las vías metabólicas que se centran en alguno o algunos de los tres elementos principales, enzimas, reacciones y compuestos. En este trabajo se presenta un análisis basado en enzimas, ya que el principal interés es la historia evolutiva de las vías, en las que intervienen directamente estas proteínas. La información sobre las reacciones y los compuestos involucrados se puede obtener de los números enzimáticos, código que se usa para identificarlas [18].

El alineamiento de vías metabólicas representa una de las herramientas más poderosas para el análisis del metabolismo. Involucra el reconocimiento de metabolitos comunes a un conjunto de vías metabólicas que pueden estar o no funcionalmente relacionadas, así como la interpretación de la evolución biológica y la determinación de vías metabólicas alternas. Además, es de gran ayuda para la predicción de funciones así como en el modelado del metabolismo [14].

Aunque la investigación en alineamiento de secuencias genómicas es extensa, el problema del alineamiento de vías metabólicas ha recibido menos atención. A continuación se describe el panorama general del problema del alineamiento de secuencias biológicas y posteriormente se hace una revisión del estado del arte en el alineamiento de vías metabólicas.

3.2 Alineamiento de secuencias biológicas

La comparación de las secuencias es utilizada para evaluar la homología (procedencia de un ancestro común) de genes y proteínas, clasificación, predicción de funciones y

estructuras (secundarias y terciarias), detectar mutaciones puntuales, construcción de árboles filogenéticos, entre otras [41].

Los eventos evolutivos que se tratan de identificar al realizar estos alineamientos son las mutaciones, inserciones y deleciones, ya que son los eventos que aportan ideas sobre los orígenes de estas secuencias. Cuando se trabaja con proteínas se tienen secuencias que contienen 20 posibles símbolos representando a los aminoácidos (unidad básica de las proteínas), y en el caso de ADN, se tienen cuatro símbolos, las cuatro diferentes bases nucleicas [32]. Al alinear las secuencias se posicionan los símbolos uno sobre otro y se agregan "*gaps*" (inserción de uno o más de estos símbolos: "-"), buscando alinear las posiciones idénticas y las similares. Si se alinean dos símbolos idénticos (que no sean "-"), entonces se supone que no existió algún cambio, si son diferentes se interpreta como una mutación puntual, y si se tiene una alineación con un "-" (*gap*), se supone que ocurrió una inserción en una de las secuencias y una deleción en otra.

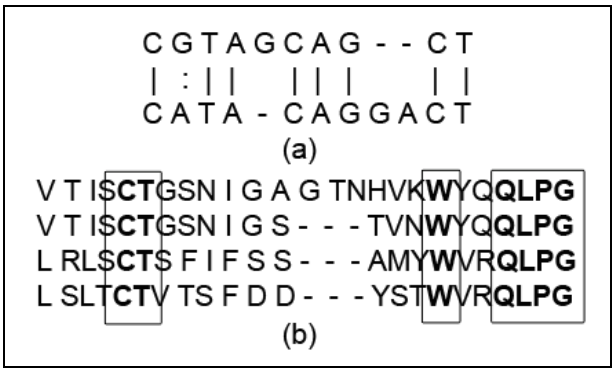


Figura 3.1. En (a) se muestra un alineamiento por pares de secuencias de ADN. El símbolo "|" representa dos elementos iguales alineados. Cuando se tienen dos elementos diferentes se utiliza ":" y cuando se alinea un elemento y un *gap* ("-") no se utiliza un símbolo. En (b) tenemos un alineamiento múltiple de cuatro secuencias de aminoácidos, las columnas marcadas con rectángulos indican los aminoácidos conservados (los que no sufrieron cambios a lo largo de la evolución). Observamos de nuevo la presencia de "-" para lograr un mejor alineamiento.

Desde esta perspectiva, el objetivo es reconstruir la historia evolutiva del grupo de secuencias insertando *gaps* para reflejar los eventos ocurridos [5]. Entonces el problema se

vuelve combinatorio, debido a la infinidad de posibilidades que se tiene para acomodar los *gaps* de forma que reflejen las mutaciones, inserciones y deleciones.

Existen dos tipos de alineamientos, los alineamientos por pares y los múltiples, los primeros pueden ser considerados como un caso especial de los alineamientos múltiples (Figura 3.1). En el primer caso se emplean principalmente técnicas de programación dinámica, que aseguran encontrar el alineamiento óptimo según determinado criterio. Sin embargo, estas técnicas no se pueden extender para emplear directamente en el segundo caso, ya que la complejidad computacional crece en el orden de $O(L^n)$, donde L es la longitud de la secuencia más larga y *n* es el número de secuencias. Para cualquiera de estos dos tipos de alineamientos, existen dos formas de realizarlos, el alineamiento local que determina subsegmentos de una secuencia dentro de otra (se utilizan principalmente para buscar homólogos dentro de bases de datos), y los alineamientos globales donde cada elemento de una secuencia es comparado con cada elemento de la otra (son usados en estudios evolutivos) [5].

Determinar cuál es el mejor alineamiento depende de la forma en que tal alineamiento es evaluado. Uno de los esquemas de evaluación más populares consiste en usar matrices de sustitución (matrices que reflejan la probabilidad de que los elementos hayan quedado alineados dado que provienen de un ancestro común, con respecto a la probabilidad de que se hayan alineado al azar). Se suelen usar en el contexto de alineamientos de aminoácidos o ADN. Las más utilizadas son las matrices PAM y BLOSUM [4,25]. Otro de los esquemas más utilizados es la suma de pares, en el cual el alineamiento es visto como una matriz de *N x M*, done *N* es el número de secuencias y *M* es la longitud de las mismas. Se califican todas las comparaciones por pares entre cada elemento de cada columna y se suman. Para evaluar las comparaciones por pares se pueden usar matrices de sustitución o algún otro esquema que determine el costo del alineamiento de dos elementos iguales, diferentes o con un *gap* [11]. Esta es la definición:

$$SumaDePares = \sum_{k=1}^{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} s(p_{ik}, p_{jk})$$

Ecuación 3.1

Donde $s(p_{ik}, p_{jk})$ es la puntuación que se le otorga al alineamiento del par de símbolos de las secuencias *i* y *j* en la columna *k*. La puntuación para el alineamiento es la suma sobre todas las columnas del alineamiento.

Existe una gran cantidad de trabajos relacionados con el problema del alineamiento de secuencias biológicas; para el problema del alineamiento por pares destacan las propuestas de programación dinámica de Needleman & Wunsh y Smith & Waterman [38,55], junto con un gran número de variaciones de ellas. En el caso de los alineamientos múltiples, debido a la complejidad del problema, se requiere de métodos más sofisticados, generalmente heurísticos, dentro de los cuales encontramos los modelos ocultos de Markov, el alineamiento progresivo (el más utilizado), métodos iterativos, el recocido simulado y el cómputo evolutivo. Junto con estos métodos también encontramos diversas herramientas que implementan estos métodos, entre las cuales están FASTA [44], BLAST [3] (alineamiento por pares), CLUSTALW [52], T-Coffe [40], MUSCLE [19] (alineamiento múltiple), entre muchas otras. En [41] encontramos una revisión de los principales métodos y herramientas disponibles para resolver el problema del alineamiento múltiple de secuencias. Sin embargo, estas herramientas han sido principalmente desarrolladas para secuencias de ADN y aminoácidos.

3.2.1 Cómputo Evolutivo y Alineamiento de Secuencias

Una de las técnicas más populares del cómputo evolutivo son los algoritmos genéticos, que como los algoritmos evolutivos en general, se inspiran en el proceso evolutivo que ocurre en la naturaleza, específicamente en la selección natural de los individuos más aptos. Holland [27] presenta el algoritmo genético (AG) como un método de optimización donde una población de “cromosomas” (por ejemplo, cadenas de 0’s y 1’s) va evolucionando a lo largo de un número de generaciones utilizando operadores genéticos como son la selección, la recombinación y la mutación. La idea principal es que las mejores soluciones son seleccionadas para ser recombinadas y crear nuevas soluciones más aptas, después éstas son mutadas con una probabilidad muy baja, con el objetivo de crear variabilidad dentro de

la población. Finalmente, después de cierto tiempo, el algoritmo encuentra la solución óptima o una muy cercana [37].

La estructura general del algoritmo genético es la siguiente:

1. *Generar la población inicial*
2. *Evaluar la población*
3. *Generar una nueva población usando los operadores (selección, mutación, cruza y otros).*
4. *Volver a 2 mientras no se cumpla la condición de paro.*

Los principales elementos del AG son la representación de las soluciones, la función objetivo y los operadores que actúan sobre los individuos de la población [37].
Una de las principales ventajas que ofrecen estos algoritmos es la posibilidad de explorar un amplio espacio de búsqueda complejo usando un grupo de soluciones a la vez. Otra es la separación entre el proceso de optimización y el criterio de evaluación (función objetivo).
Adicionalmente, este tipo de algoritmos han sido utilizados en diversos ámbitos como problemas de negocios, aplicaciones en la ciencia y en la ingeniería [12,57] y la investigación en la teoría subyacente es un campo de gran interés, donde constantemente se está en busca de un consenso para explicar cómo es que funcionan.
Las técnicas de cómputo evolutivo que han sido aplicadas al problema del alineamiento de secuencias incluyen a la programación evolutiva [10] y a los algoritmos genéticos [15, 23, 29]. También podemos encontrar algoritmos híbridos que combinan dos o más de estas técnicas, como en [31], donde se propone un algoritmo genético combinado con la optimización basada en colonias de hormigas para resolver el problema de alineamiento múltiple y en [42] donde se usa un algoritmo genético y el recocido simulado. Dentro de los AGs tenemos el trabajo de Arenas, que propone un algoritmo genético para mejorar el alineamiento múltiple de secuencias génicas y de proteínas [6]. En estos trabajos encontramos diferentes formas de representar los alineamientos dentro del AG, nuevos operadores de recombinación y mutación, dependientes de la representación, así como distintas propuestas de evaluación.

3.3 Alineamiento de vías metabólicas

Las herramientas existentes para la comparación de vías metabólicas son relativamente pocas, y muy diferentes, ya que como se mencionó anteriormente, dependiendo del enfoque del análisis que se quiera realizar se utilizan diferentes abstracciones. Podemos resumir los enfoques que se proponen en dos clases, los que se basan en un análisis de grafos, tomando en cuenta la topología de las vías y ciertas propiedades que se pueden derivar de ella, y los algoritmos que no le dan importancia a la topología, centrándose principalmente en los elementos de las vías, ya sean las enzimas, compuestos y/o reacciones. Existen también algunos trabajos que combinan estos dos enfoques.

En la Figura 3.3 se muestra una de las vías metabólicas más conocidas, la de la glucólisis. Esta es la representación que se utiliza en la base de datos KEGG[2] (*Kyoto Encyclopedia of Genes and Genomes*), donde se manejan mapas de referencia (vías comunes a todos los organismos que se encuentran reportados en la base) y de acuerdo al organismo que se desea analizar, las enzimas específicas son resaltadas usando un color diferente. Se pueden observar los tres elementos principales mencionados anteriormente.

En la Figura 3.4 podemos observar un ejemplo de la abstracción de una vía para ser representada como un grafo, donde los nodos pueden ser cualquiera de los tres elementos principales de la vía.

3.3.1 Algoritmos basados en la topología de las vías

La mayoría de los trabajos recientes para comparar vías se encuentran en el problema de comparación de redes de interacción de proteínas (la unión de dos o más proteínas para llevar a cabo su función biológica). Algunos de estos trabajos se pueden extender a las vías metabólicas. Generalmente utilizan un modelo de grafo para representar las vías, donde los nodos son enzimas o compuestos y las aristas son las reacciones, entonces el problema es equivalente al problema del isomorfismo de grafos/subgrafos, y dado que este problema es

[2] Disponible en http://www.genome.jp/kegg/

NP-completo, el problema del alineamiento de las vías usando esta representación también lo es, por lo que se requieren heurísticas para resolverlo [7,14].

Los métodos que emplean grafos para la representación generalmente usan ciertas restricciones para las topologías de las vías. Algunos modelos que se han empleado para la representación son las redes lógicas y booleanas, en los que se usan los vértices para denotar enzimas y compuestos, y las interacciones entre estos son representadas por las aristas. Consideran un nodo como un operador "AND" ú "OR" y asignan valores booleanos a las aristas. Otros modelos comúnmente usados para representar interacciones de este tipo son las redes bayesianas, los hiper-grafos y redes de Markov [7].

En muchos de estos trabajos se trata de capturar ciertas propiedades presentes en los grafos que representan a estas redes biológicas y usarlas como medidas de similitud. En otros casos se presentan patrones frecuentes que se encuentran dentro de las redes [46].

Dentro de las herramientas existentes que utilizan el enfoque de la teoría de grafos encontramos *MetaPathwayHunter* [45], que es una herramienta para el alineamiento de vías, la cual recibe de entrada una vía "pregunta" y utiliza una colección de vías como base de datos. Busca y reporta todas las ocurrencias aproximadas de la "pregunta" dentro de la colección, ordenadas por similitud y significancia estadística. Está basado en un algoritmo de alineamiento de grafos que utiliza el modelo de homeomorfismo[3]. Los nodos representan las enzimas involucradas en la vía, y se les asigna etiquetas correspondientes a los números enzimáticos. Después se construye una tabla de sustitución que es usada para calificar los alineamientos. La evaluación de los alineamientos de los grafos es similar a la utilizada en los alineamientos de secuencias, donde se usan calificaciones para la sustitución de nodos, el borrado y la inserción de indeles, este último con una penalización fija. También ofrece una interface visual que despliega el alineamiento de dos vías homólogas. En la Figura 3.2 se muestra un ejemplo del alineamiento de salida que se obtiene con esta herramienta.

Berg et al [9] utiliza la técnica de recocido simulado para alinear dos redes de interacción de proteínas, usando ciertos patrones de familias de proteínas similares pero no son

[3]En teoría de grafos, se dice que dos grafos G_1 y G_2 son homeomorfos si ambos pueden obtenerse a partir de un mismo grafo por una sucesión de subdivisiones elementales de aristas.

patrones necesariamente idénticos. Proponen un algoritmo para el alineamiento local de grafos, que es conceptualmente similar al alineamiento de secuencias, donde resuelven el problema creando un mapeo del alineamiento de grafos al modelo de *spines* conocido dentro de la física estadística, el cual es atacado usando el recocido simulado.

Este tipo de representaciones es más adecuado cuando en el análisis es de interés la topología de la red, lo cual entonces se transforma en el problema de comparación de redes biológicas, en el que una estructura lineal ya no sería apropiada [14].

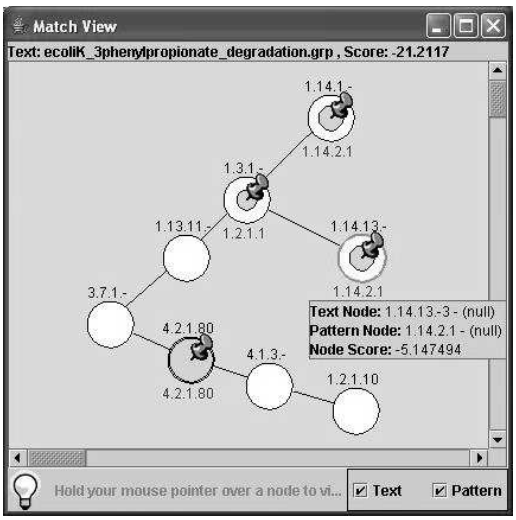


Figura 3.2. Los nodos blancos representan la vía que se desea comparar con la base de datos. Los círculos grises son la vía que obtuvo un mejor alineamiento dentro de la base de datos. Cuando se alinean dos enzimas idénticas el círculo gris cubre por completo al blanco.

3.3.2 Algoritmos basados en la composición de las vías

Dentro de los primeros trabajos en el alineamiento de vías metabólicas encontramos el de Dandekar [16], que combina varios tipos de información para la comparación de vías bioquímicas, centrando su análisis en la vía de la glucólisis para distintos organismos. En este método utiliza la información bioquímica de la vía para el análisis de los flujos metabólicos de los sustratos y un análisis comparativo de genomas, basado principalmente en las enzimas que participan en la vía, usando los genes que codifican estas enzimas. Forst y Schulten [21], extendieron los métodos de alineamiento de secuencias de ADN para definir distancias entre vías metabólicas combinando la información de las secuencias de

los genes involucrados. Usaron esta información junto con la información de la vía correspondiente, incluida la topología del grafo que la representa. Para cada rol funcional de la vía, todos los genes dentro del genoma que codifican para ese rol específico fueron utilizados. El objetivo fue realizar un análisis evolutivo de las vías metabólicas dentro de distintos organismos.

En el 2000, Tohsato [54] propone un algoritmo para el alineamiento múltiple de vías metabólicas basándose en las similitudes entre las reacciones involucradas, representadas por las similitudes entre los números enzimáticos de las enzimas respectivas. Para realizar la comparación de las vías utiliza la jerarquía en la que se encuentran clasificados los números enzimáticos (cuatro niveles), considerando que la relación entre la proximidad dentro de la jerarquía enzimática y la similitud de las reacciones es fuerte. El alineamiento por pares usa una variación del algoritmo de Needleman y para evaluar el alineamiento introduce una fórmula que calcula el contenido de información del mismo. La extensión para el alineamiento múltiple sigue el enfoque de los alineamientos progresivos, es decir va agregando una secuencia (vía) a la vez y realiza un alineamiento por pares con el resultado anterior y la nueva vía, hasta haber alineado todas.

Liao [34], desarrolló un método computacional para comparar organismos, basado en un análisis de todas las vías metabólicas. La presencia o ausencia de las vías en los organismos fue presentada como un vector booleano. Basado en esta metodología definió ciertos “perfiles” de los organismos para compararlos por pares usando una medida de distancia específica.

En [14] se presenta un algoritmo para el alineamiento lineal de vías metabólicas, un enfoque diferente a los presentados anteriormente. Los autores consideran las vías metabólicas no como un camino bien definido, ya que muchas veces contienen ramificaciones o vías alternas, lo cual origina que podamos encontrar varias vías dentro de una sola. Definen una vía metabólica como un subconjunto de reacciones enzimáticas sucesivas, donde cada reacción es catalizada por una enzima específica representada por el número enzimático único. Así, las vías están representadas por una o varias secuencias de números enzimáticos.

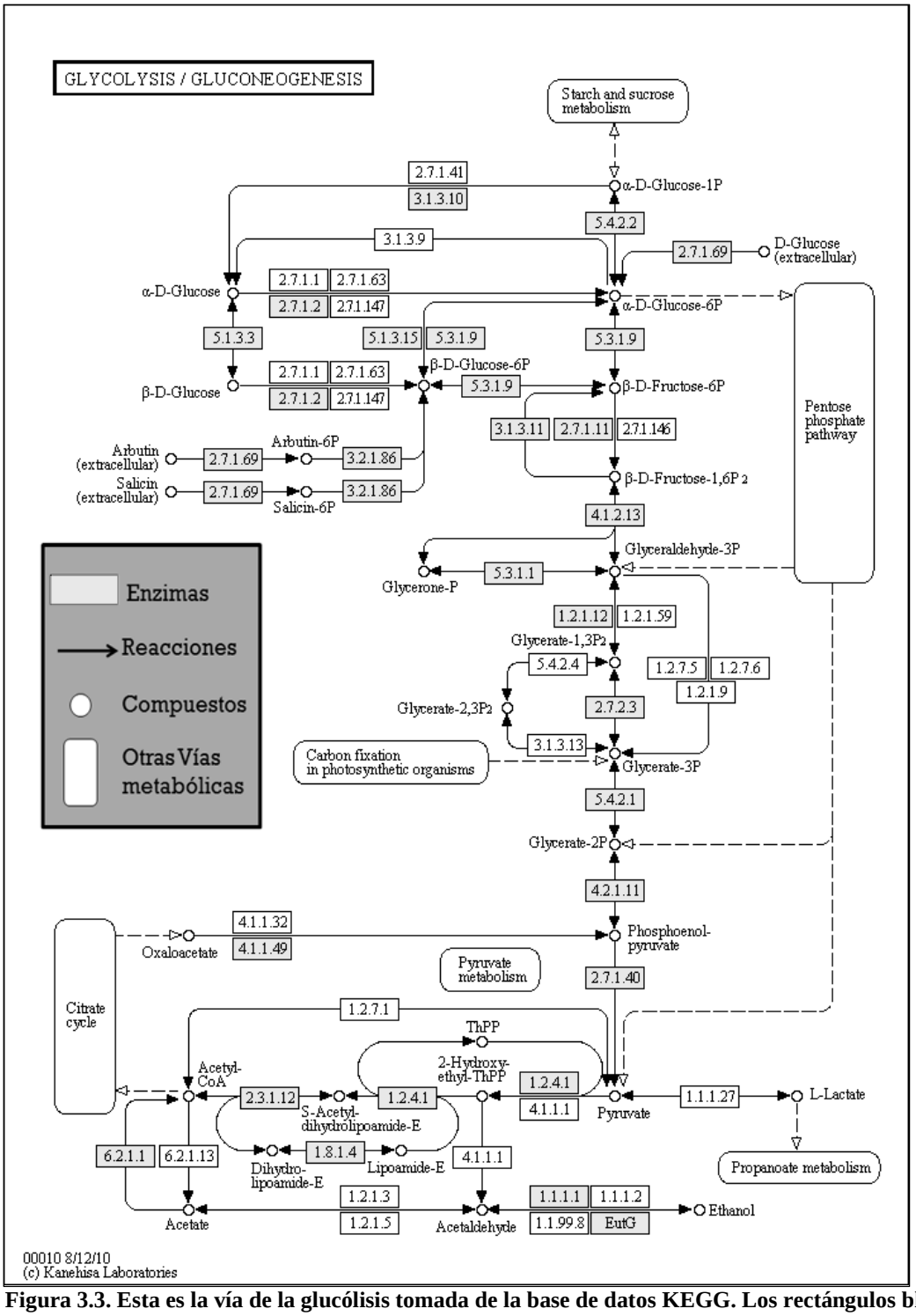


Figura 3.3. Esta es la vía de la glucólisis tomada de la base de datos KEGG. Los rectángulos blancos representan las enzimas del mapa de referencia, y los grises son las enzimas específicas para el metabolismo de *Escherichia coli*.

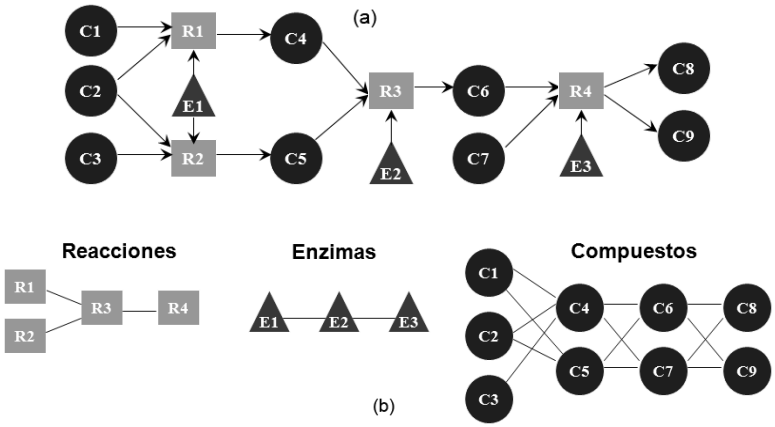


Figura 3.4. Representación de una vía metabólica usando grafos. En (a) observamos la representación de la vía con los tres elementos compuestos (C), reacciones (R) y enzimas (E). En (b) tenemos la representación de la vía usando solamente uno de los elementos a la vez.

En cuanto al alineamiento y su evaluación, se definen tres operaciones basándose en el alineamiento de cadenas de secuencias, el borrado de una enzima, la inserción y el reemplazo de una enzima por otra. Cada una de estas operaciones recibe un valor real positivo. En el caso de las operaciones de reemplazo, es decir, cuando una enzima es alineada con una que difiere a ella en cualquiera de los cuatro posibles niveles, una función de similitud es definida asignando un mayor peso a la similitud en los primeros niveles. El algoritmo propuesto puede ser extendido para realizar alineamientos múltiples, la idea general es construir una sucesión de alineamientos por pares.

PathAligner[4] es la implementación de este método. Esta herramienta se encuentra disponible en la red, donde es posible comparar dos o más vías metabólicas para evaluar su similitud, así como obtener una representación gráfica del resultado del alineamiento. En la Figura 3.5 se muestra un ejemplo de la salida que se obtiene con esta herramienta.

Otro trabajo donde la topología de la red metabólica no es importante lo encontramos en [15], donde se presenta un método para alinear vías metabólicas, basado en la jerarquía de enzimas. Usando la información que se encuentra disponible en la base de datos KEGG, proponen usar el método de alineamiento para encontrar similitudes entre las reacciones de distintos organismos, la diferencia que se muestra en este método es que no se ven las reacciones como pasos consecutivos, pero se busca la presencia o ausencia de éstas, y

[4] http://bibiserv.techfak.uni-bielefeld.de/pathaligner/

definen tres tipos posibles de alineamientos entre reacciones, el alineamiento perfecto, la sustitución y el alineamiento con *gaps*. El primero se refiere a dos reacciones en las cuales todos los elementos (sustrato, enzima y producto) son los mismos, el segundo se utiliza cuando algún elemento es diferente, y el último ocurre cuando las dos reacciones son completamente diferentes en los dos organismos que son comparados.

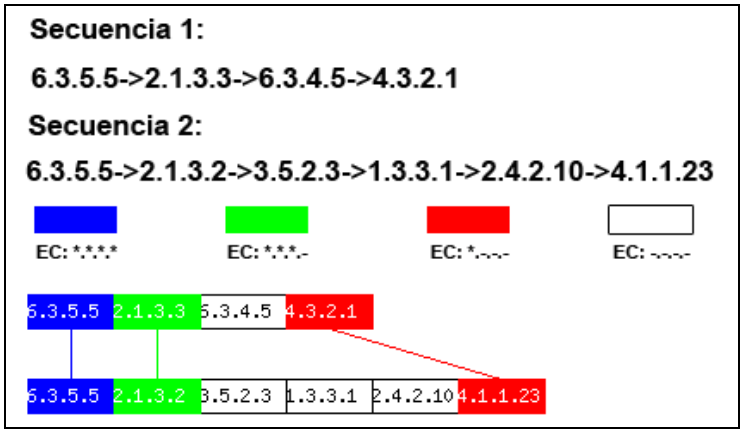


Figura 3.5. La secuencia 1 y 2 se alinean usando *PathAligner*. Cada rectángulo representa una enzima, podemos ver el número enzimático (EC) de cuatro niveles. Se usa un código de colores para indicar cuáles son los niveles de los EC que fueron iguales en cada enzima.

SIGNALIGN[5] es una herramienta web para el alineamiento de vías bioquímicas, metabólicas, de señalización y regulatorias. En este trabajo se propone un enfoque basado en el uso de la información estructural de las proteínas que intervienen en la vía, que es definida como un conjunto o una lista ordenada de proteínas que interactúan y además comparten características similares de estructura y dominio con sus contrapartes. La información es recopilada de varias bases de datos y compilada para aplicar ciertas medidas de similitud al realizar el alineamiento. Para evaluar el alineamiento se mide el contenido de información que se comparte entre los elementos que son alineados, además de incluir una penalización por inserción de indeles [24].

En [33] se propone un método para alinear redes metabólicas completas y cuantificar sus similitudes a fin de encontrar vías altamente conservadas en distintos organismos. Aquí se define una vía como una serie de reacciones químicas del metabolismo dentro de una célula, así que no son vistas necesariamente como rutas dentro de la red. En lugar de aislar

[5]http://agbi.techfak.uni-bielefeld.de/signalign/index.jsp

las reacciones siguiendo una ruta de entrada y salida, se buscan vías conservadas, y no reacciones. La idea es ir creando bloques constructores a medida que se van encontrando similitudes entre las vías que se están alineando, se definen varios tipos de bloques dependiendo del nivel de similitud que exista entre las entidades que se comparan.
La función de evaluación propuesta en este trabajo integra información de los sustratos, productos, la funcionalidad de las enzimas así como sus secuencias, además de la topología de los alineamientos. Esta función logra reflejar una gran cantidad de información del problema biológico que se trata de resolver.
En [53] Tohsato hace una nueva propuesta para alinear vías metabólicas. Este nuevo enfoque se basa en las similitudes entre las estructuras químicas de los compuestos que intervienen en las vías. El algoritmo de alineamiento es una variación del método basado en la programación dinámica, tratando de alinear reacciones similares, evaluando esta similitud con las estructuras de los compuestos involucrados.

Dentro de los métodos que combinan los dos enfoques, recientemente Ferhat [7] propone un algoritmo para el alineamiento por pares utilizando un modelo de grafos para la representación de las vías metabólicas tomando en cuenta las reacciones, compuestos y enzimas. Un problema de *eigen*-valores es creado para cada entidad a fin de evaluar las similitudes y realizar un mapeo entre las vías usando el *"matching" del máximo grafo bipartito*. Al usar este método de la teoría de grafos e incluir a las tres entidades en la comparación, la topología de la vía tiene un gran peso al momento de realizar el alineamiento. También proponen una medida de similitud para evaluar los alineamientos obtenidos y ofrecen pruebas estadísticas de la eficiencia de la función de similitud. Cabe mencionar que el modelo que proponen toma en cuenta las tres entidades, pero tiene la opción de restringir la comparación a una sola de ellas, como las enzimas, para realizar otro tipo de análisis. El código fuente se encuentra disponible en la página web del autor[6].
En varios de los métodos mencionados al realizar los alineamientos no resulta tan fácil observar gráficamente las similitudes y diferencias cuando se trata de más de dos vías, por lo cual se requiere de otro tipo de representación para lograr esto.

[6] http://bioinformatics.cise.ufl.edu/palAlign.html

Podemos observar que el problema tiene tres componentes principales, el modelo adecuado para representar las vías metabólicas dependiendo de lo que se quiera analizar, el algoritmo a emplear para realizar el alineamiento y finalmente el criterio que se utiliza para evaluar los alineamientos. Otros puntos importantes que considerar son la base de datos que se utiliza y como mostrar el resultado obtenido para que sea fácil de analizar visualmente.

En el siguiente capítulo se describe la metodología de este trabajo para resolver el problema del alineamiento de vías metabólicas.

Capítulo 4

METODOLOGÍA

La herramienta principal para realizar el análisis comparativo de las vías metabólicas es el algoritmo que nos permite crear los alineamientos múltiples. En este capítulo se describe la metodología propuesta para llevar a cabo este análisis, incluido el algoritmo de alineamiento y otras herramientas que se proponen para mejorar la calidad del análisis.

4.1 Base de datos

Actualmente están disponibles diversas bases de datos que incluyen la información de vías metabólicas conocidas. En la Tabla 4.1 se muestran algunas de dichas bases de datos [13].

Base de datos	Sitio web	Descripción
KEGG	http://www.kegg.jp/kegg/pathway.html	Contiene representaciones gráficas de todas las vías metabólicas conocidas de varios organismos.
EcoCyc	http://ecocyc.org/	Es una base de datos para la bacteria *Escherichia coli K-12.* Contiene información del genoma, de las redes de transcripción y regulación, además de las vías metabólicas.
MetaCyc	http://metacyc.org/	Contiene información de vías metabólicas de varios organismos obtenidas experimentalmente.
UniPathway	http://www.grenoble.prabi.fr/obiwarehouse/unipathway	Es una colección de vías metabólicas depuradas manualmente para ser usadas junto con la base de datos UniProtKB/Swiss-Prot.

Tabla 4.1. Bases de datos de vías metabólicas disponibles en la red.

La base de datos *Kyoto Encyclopedia of Genes and Genomes* (KEGG) fue seleccionada para la realización de este trabajo debido a que el formato que se emplea es fácil de manejar, además de ser una de las que contiene una información de mejor calidad acerca de metabolismo celular. La información se encuentra almacenada en archivos XML (*eXtensible Markup Language*) a partir de los cuales se crean representaciones gráficas de las vías metabólicas. Consiste de tres tipos principales de datos que se encuentran en las siguientes bases:

1. PATHWAY: Integra el conocimiento sobre las redes de interacción molecular, tales como las vías metabólicas.
2. GENES/SSDB/KO: Información acerca de los genes y las proteínas.
3. COMPOUND/GLYCAN/REACTION: Información sobre compuestos y reacciones bioquímicas.

Adicionalmente, KEGG maneja mapas de referencia para cada una de las vías metabólicas conocidas, y a partir de ellos se generan mapas para los organismos específicos que se encuentran reportados en dicha base de datos [28].

La información de la bacteria *Escherichia coli K-12* se eligió para ser el caso de estudio, ya que es uno de los organismos que contiene la mayor cantidad de información biológica bien determinada.

En total se trabajó con 47 vías metabólicas, que contienen 384 enzimas diferentes representadas por 86 diferentes números enzimáticos (3 niveles). El número total de reacciones es de 861. Los archivos que contienen los mapas se tomaron de la versión de KEGG liberada el 30 de septiembre de 2009, disponible en la siguiente dirección:

ftp://ftp.genome.jp/pub/kegg/release/archive/kegg/52/

Los archivos XML se encuentran en la ruta: *XML/organisms/eco.* El archivo *eco_enzime.lst* contiene la información de los números enzimáticos.

Las vías metabólicas de *E. coli* fueron transformadas a la representación propuesta de la siguiente manera.

4.2 Representación de las vías metabólicas

Anteriormente se mencionó que existen varios trabajos en los que se representan las vías metabólicas por medio de secuencias de números enzimáticos. En [36] se sigue esta idea, construyendo varios grafos a partir de los mapas de KEGG y después recorriéndolos para obtener las secuencias. Se retomó esta idea realizando algunas modificaciones para la obtención de las secuencias.

En la Figura 3.3 tenemos uno de los mapas que se encuentran en KEGG; toda la información que aparece en ellos se encuentra en los archivos XML, que son los que se utilizaron para la reconstrucción. El objetivo es crear un árbol donde los nodos son las enzimas que catalizan las reacciones de la vía y las aristas representan estas reacciones.

Dentro de la base de datos, las enzimas tienen como atributos un identificador único, el número enzimático correspondiente y las reacciones en las que intervienen, que pueden ser una o más. Debido a que la clasificación en el cuarto nivel de los números enzimáticos es ambigua o muchas veces no existe, se decidió utilizar únicamente los primeros tres niveles para representar a las enzimas.

Las reacciones tienen asociadas además de un identificador único, el sustrato y el producto que intervienen en ellas. Estos atributos nos sirven para crear una lista de adyacencia[7] de las enzimas de la vía, que serán los nodos del grafo que se reconstruye. *Dos enzimas E1 y E2, son adyacentes si E1 tiene una reacción asociada, cuyo producto es el sustrato de una segunda reacción, asociada a E2.*

Después de crear la lista de adyacencia, el siguiente paso es seleccionar los nodos de inicio para recorrer el árbol. Ya que esta información no se encuentra disponible en KEGG, definimos un *nodo de inicio* como la enzima que catalice una *reacción cuyo sustrato no sea producto de ninguna otra reacción*.

Para cada vía tenemos varios posibles nodos de inicio, cada uno de ellos es la raíz de un árbol. Al recorrer cada uno de estos árboles desde la raíz hasta las hojas, obtenemos las secuencias de enzimas que representan las vías. El recorrido se hace con el algoritmo "*Breadth First Search*" (BFS).

[7] Forma de representar un grafo, donde cada vértice tiene una lista de los vértices adyacentes a él.

4.2.1 Breadth First Search

El algoritmo de "búsqueda a lo ancho" es un algoritmo exhaustivo usado para recorrer o buscar datos en un grafo, generalmente representado como un árbol. Se comienza en la raíz y se exploran todos los vecinos de este nodo. Después, para cada uno de los vecinos se exploran sus respectivos vecinos adyacentes, y así hasta que se recorra todo el árbol. Siempre se elige el nodo menos profundo para ser expandido, esto se logra usando una cola tipo FIFO (*First In, First Out*), donde el primer elemento en entrar a la cola es el primero en salir (Figura 4.1) [48]. De esta forma, se logra obtener todos los caminos posibles desde cada uno de los nodos de inicio de la vía, incluyendo las ramificaciones que se pudieran encontrar. En la Figura 4.2 se muestra un ejemplo del proceso de reconstrucción.

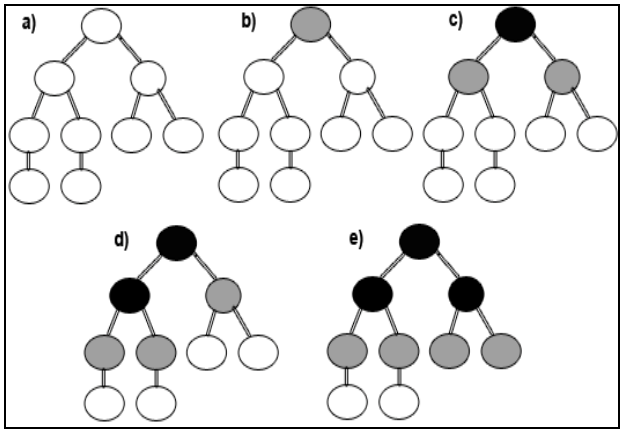


Figura 4.1. Ejemplo del recorrido del árbol con el algoritmo BFS. El recorrido se realiza primero a lo ancho y después a lo profundo. De color gris aparecen los nodos que se encuentran en la pila y en negro los nodos que ya fueron procesados. Los nodos de color blanco son los que faltan por procesar.

Después de aplicar este algoritmo a los 47 mapas de *E. coli*, obtenemos 452 secuencias de enzimas que representan estas vías metabólicas. El número de secuencias correspondientes a cada vía se encuentra en el apéndice A.

4.3 Algoritmo Genético

Un algoritmo genético fue implementado para realizar los alineamientos múltiples de las secuencias que se obtuvieron. A continuación se describen los principales elementos de este algoritmo.

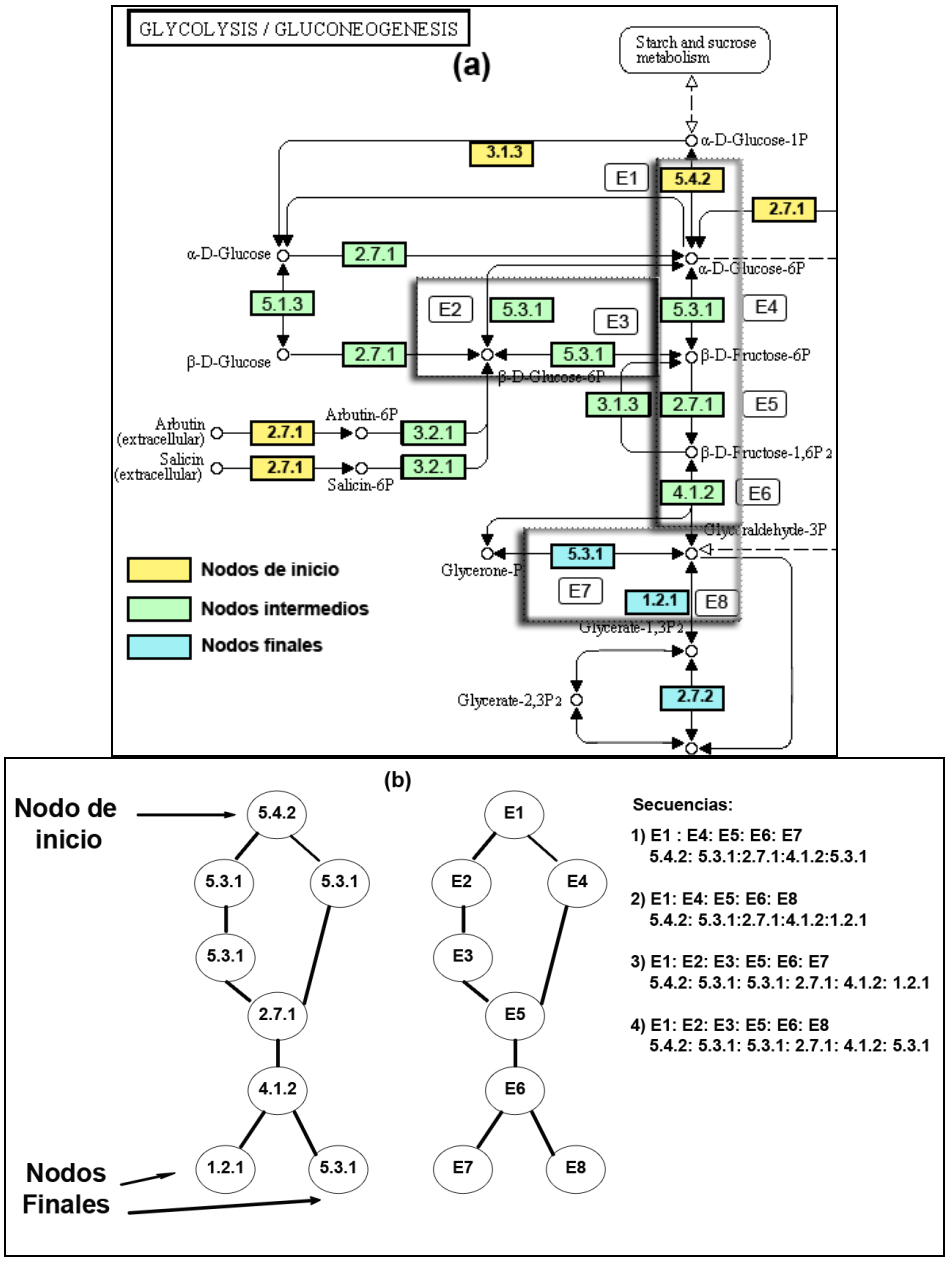


Figura 4.2. (a) Segmento de la ruta de la glucólisis. Los rectángulos representan a las enzimas (E1, E2, E3, E4, E5, E6, E7 y E8), los círculos son los sustratos y productos. De acuerdo al algoritmo de reconstrucción de secuencias E1 es un nodo de inicio. Siguiendo la secuencia de reacciones obtenemos las listas de adyacencia de las enzimas y en (b) tenemos el árbol que se forma. Finalmente se tienen cuatro secuencias, resultado de la ejecución del algoritmo BFS.

4.3.1 Representación

Los individuos de la población representan posibles alineamientos múltiples en forma de una matriz, donde cada fila corresponde a una secuencia y cada columna a una posición en el alineamiento. Dado que la longitud de las secuencias en el alineamiento puede ser diferente, al final de cada secuencia se insertan *gaps* ("-") para que todas alcancen la longitud de la secuencia más larga, es decir, para homogeneizar el alineamiento. Este es un problema intrínseco de la representación empleada.

Los símbolos que se alinean son los números enzimáticos (tres niveles), en donde en cada columna hay un número que tiene esta forma: A.B.C. En la Tabla 4.2 se muestra un ejemplo.

	Sin alinear					**Alineamiento**					
# secuencia	**0**	**1**	**2**	**3**	**4**	**0**	**1**	**2**	**3**	**4**	**5**
1	1.1.1	1.2.3	2.7.9	1.2.2	6.2.1	1.1.1	1.2.3	2.7.9	1.2.2	6.2.1	-.-.-
2	1.1.1	1.2.2	6.2.1	3.2.1		1.1.1	-.-.-	-.-.-	1.2.2	6.2.1	3.2.1
3	1.1.7	1.2.2	6.2.1	3.2.1		1.1.7	-.-.-	-.-.-	1.2.2	6.2.1	3.2.1
4	1.2.1	2.7.2	6.2.1			-.-.-	1.2.1	2.7.2	-.-.-	6.2.1	-.-.-

Tabla 4.2. A la izquierda se presentan cuatro secuencias sin alinear. A la derecha se muestra la matriz de un alineamiento generado (representación original, que se obtiene al decodificar la representación en el AG). Debido a la inserción de *gaps*, la longitud de las secuencias aumenta a un total de 5 columnas.

A partir de esta matriz se realizó la codificación para el algoritmo genético. Se utiliza una codificación binaria, donde las enzimas (números enzimáticos) son representados por 1's y los 0's representan los *gaps* que son insertados para maximizar el número de columnas alineadas. Los *gaps* son insertados entre enzimas, no entre los números que representan la clasificación de una enzima, es decir, no se puede insertar un "-" para que la enzima quede de esta forma: A.-.C.
El cromosoma del individuo presentado en la Tabla 4.2, se muestra en la Tabla 4.3.

# secuencia	0	1	2	3	4	5
1	1	1	1	1	1	0
2	1	0	0	1	1	1
3	1	0	0	1	1	1
4	0	1	1	0	1	0

Tabla 4.3. Representación (codificación) en el AG. El "1" representa las enzimas y los ceros a los *gaps* (-.-.-).

4.3.2 Operadores

Los operadores del AG nos permiten explorar el amplio espacio de búsqueda de soluciones del problema. En la Figura 4.3 se muestra el esquema básico del AG.

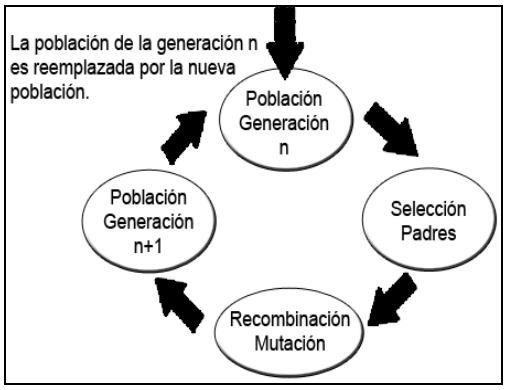


Figura 4.3. Esquema básico de la evolución de la población en el AG.

4.3.2.1 Selección

Es el método que servirá para seleccionar a los individuos de la población que darán origen a la siguiente generación. También define cuántos descendientes tendrá cada uno de los individuos. El propósito es preferir a los individuos con mayor aptitud dentro de la población esperando que su descendencia tenga una aptitud aún mayor. Existen varios esquemas de selección, dentro de los más comunes se encuentran el método por rango, la selección proporcional, la selección elitista, la selección por ruleta y la selección por torneo [37].

Selección por torneo

Después de varias pruebas se decidió utilizar este método de selección combinado con el elitismo (el mejor individuo de la población es copiado sin modificación a la siguiente generación). Este método de selección tiene como parámetro el tamaño del torneo, que indica el número de individuos que son elegidos al azar dentro de la población para competir entre ellos. En este caso el tamaño del torneo fue dos. El individuo más apto del subgrupo es el ganador, el que se elige para reproducirse. Esta técnica tiene la ventaja de que permite un cierto grado de elitismo, el mejor nunca va a morir, y los mejores tienen más probabilidad de reproducirse que los peores, pero sin producir una convergencia prematura, es decir, que toda la población esté compuesta del mismo individuo. Dentro de las ventajas que ofrece esta técnica están: su eficiencia para ser codificado, fácil paralelización y permite que la presión de selección (el grado en el cual los mejores individuos son favorecidos) se ajuste fácilmente, cambiando el tamaño del torneo [37].

El resultado de este operador es la población de padres que serán elegidos para crear la nueva población que reemplazará a la anterior.

4.3.2.2 Cruza

El operador de cruza o recombinación, consiste en crear nuevos individuos a partir de dos padres elegidos al azar de la población seleccionada para reproducirse, combinando los cromosomas que codifican características de las soluciones. Este operador es el responsable de explotar las soluciones para encontrar mejores. La forma de aplicarse depende mucho de la codificación de las soluciones. En el caso más común donde se utiliza una cadena de 0's y 1's, el operador de cruza elige un punto al azar dentro de la cadena para realizar el intercambio de los segmentos que fueron divididos, a fin de crear dos nuevos individuos [37].

En el caso del algoritmo propuesto se probaron varios operadores de recombinación que se describen a continuación.

Cruza – Intercambio

Recordemos que el cromosoma de los individuos está codificado como una matriz de 0's y 1's, donde las filas representan a las secuencias que se quieren alinear. En este operador se elige el punto de cruza horizontalmente, es decir, se cruzan secuencias. Se elige aleatoriamente el punto k para realizar la cruza, entonces la matriz del padre 1 (P_1) queda dividida en dos partes: M_{11} y M_{21},

donde M_{11} va de la secuencia 1 hasta la *k*, mientras que M_{21} va de la secuencia *k+1* hasta la *n* (número total de secuencias). Lo mismo sucede con la matriz del padre 2 (P_2). Al realizar el intercambio, los dos hijos se forman de la siguiente manera:

$$P_1 = \begin{bmatrix} M_{11} \\ M_{21} \end{bmatrix},\ P2 = \begin{bmatrix} M'_{12} \\ M'_{22} \end{bmatrix}$$

Ecuación 4.1

$$H_1 = \begin{bmatrix} M_{11} \\ M'_{22} \end{bmatrix}, H_2 = \begin{bmatrix} M'_{12} \\ M_{21} \end{bmatrix}$$

Ecuación 4.2

Cruza – Columnas 1

Cuando se trata de realizar la cruza usando las columnas, surge un problema. Debido a que las secuencias tienen un número fijo de enzimas, el número de 1's en cada fila de la matriz igual es fijo, entonces, al seleccionar aleatoriamente el punto de corte en las matrices de los padres, y realizar la recombinación, podemos obtener filas con un número diferente de 1's de los necesarios. Para solucionar este problema, la primera estrategia que se implementó fue realizar la recombinación normal, aunque el número de 1's que se tuviera en cada fila de la matriz fuera incorrecto. Con estas matrices, se procedió a realizar la corrección adecuada (eliminar o agregar símbolos).

Se recorre cada fila de la matriz y se cuenta el número de 1's, si es mayor al permitido, entonces se elige al azar una posición y cuando el símbolo es un uno, se cambia por un cero; esto se repite hasta tener el número correcto de 1's. Si el número es menor, se sigue el mismo procedimiento pero eligiendo posiciones que sean ceros, para ser cambiadas por unos, hasta alcanzar el número necesario. Este método resultó ser muy ineficiente debido a que la estrategia que se implementó para la corrección de los individuos es muy lenta, ya que tiene que hacer varios recorridos de toda la matriz para realizar el conteo y después para buscar las posiciones que se puedan modificar. Por estas razones se intentó una nueva estrategia.

Cruza – Columnas 2

Tenemos las matrices M_1 y M_2 de los padres, elegimos el punto *k* aleatoriamente, que nos indica la columna a partir de la cual se va a realizar la recombinación. El hijo 1 (H_1) obtiene la primera parte de su cromosoma del padre 1, de la columna *1* hasta la columna *k* (M_{11}). Para evitar agregar más 1's de los permitidos en las columnas de *k+1* hasta *m* (longitud de la secuencia más larga), se recorre el

cromosoma de la matriz M_2 de la siguiente forma: para cada fila de la matriz H_1 se cuenta el número de 1's que presenta, si ya se tiene el máximo, se agregan ceros hasta alcanzar la longitud máxima, si hacen falta, se recorre la fila correspondiente de M_2 y se van agregando en orden los elementos de esta fila hasta alcanzar el número de 1's y 0's permitidos, de forma que si el número de 0's ya fue alcanzado pero aún faltan 1's, ya no se permite agregar más y estos elementos de M_2 no son agregados. Para H_2 el proceso es similar, pero intercambiando el orden de los padres. La estructura final es la siguiente:

$$H_1 = [M_{11} \quad M_{2^*}], H_2 = [M_{21} \quad M_{1^*}]$$
$$donde,$$
$$M_1 = [M_{11} \quad M_{12}], la\ matriz\ del\ padre\ 1$$
$$M_2 = [M_{21} \quad M_{22}], la\ matriz\ del\ padre\ 2$$

Ecuación 4.3

$$M_{11} = \begin{bmatrix} m_{1,1} & \dots & m_{1,k} \\ \dots & \dots & \dots \\ m_{n,1} & \dots & m_{n,k} \end{bmatrix}, M_{12} = \begin{bmatrix} m_{1,k+1} & \dots & m_{1,m} \\ \dots & \dots & \dots \\ m_{n,k+1} & \dots & m_{n,m} \end{bmatrix}$$
$$M_{21} = \begin{bmatrix} m_{1,1} & \dots & m_{1,k} \\ \dots & \dots & \dots \\ m_{n,1} & \dots & m_{n,k} \end{bmatrix}, M_{22} = \begin{bmatrix} m_{1,k+1} & \dots & m_{1,m} \\ \dots & \dots & \dots \\ m_{n,k+1} & \dots & m_{n,m} \end{bmatrix}$$

M_{2^*} *es igual a* M_{22} *y* M_{1^*} *es igual a* M_{12} *con las* mod*ificaciones descritas.*

k es la columna elegida aleatoriamente como punto de cruza.

n *y m son las dimensiones de las matrices de los cromosomas.*

Ecuación 4.4

4.3.2.3 Mutación

Este operador previene que el algoritmo se quede atrapado en mínimos locales. Es el responsable de la exploración de todo el espacio de búsqueda, mantiene la diversidad en la población al introducir pequeñas alteraciones en los cromosomas de los individuos. Una de las formas de mutación más comunes consiste en intercambiar el valor 0 por el 1, y viceversa, en el caso de codificaciones binarias [37].

Se aplica sobre todos los *bits* del cromosoma de los individuos, con una probabilidad muy baja, determinada como parámetro del algoritmo, es decir, todos los *bits* tienen la posibilidad de ser mutados, pero el número total de mutaciones esperadas en la codificación de la matriz de

alineamiento es muy bajo, con el objetivo de no destruir soluciones buenas. Se describen los operadores de mutación que se implementaron.

Mutación – *Gaps*

Si el *bit* a mutar es un cero, entonces el *gap* puede ser extendido o bien reducido en una unidad, con igual probabilidad, es decir se agrega un cero después del bit mutado o bien se elimina el cero. Si el *bit* es un uno, se inserta un *gap* de longitud uno, es decir, insertamos un *gap* (un cero dentro de la secuencia de *bits* después del bit a mutar). Este operador implica la posterior corrección de la matriz agregando ceros a las filas para que se tengan la misma longitud en todas ellas.

Mutación – Circular

Si el *bit* que se está evaluando es elegido para ser mutado, la modificación consiste en realizar un corrimiento circular hacia la derecha de la fila en la que se encuentra, a partir de la columna del *bit* a mutar. Entonces, es posible que se realicen varios corrimientos dentro de la matriz completa, que es equivalente a cambiar las posiciones donde se encuentran los *gaps*, sin aumentar el tamaño. En la Tabla 4.4 se muestra un ejemplo.

	Antes de la mutación						**Después de la mutación**					
# secuencia	**0**	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**
1	1	1	1	1	1	0	1	1	1	1	1	0
2	1	0	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**
3	1	0	0	1	1	1	1	0	0	1	1	1
4	0	1	1	0	1	0	0	1	1	0	1	0

Tabla 4.4. El *bit* 2 de la secuencia 2 es elegido para aplicar el corrimiento a la derecha.

4.3.3 Evaluación de los alineamientos

Para realizar la evaluación de los alineamientos, es necesario realizar una transformación para regresar a las enzimas y *gaps* que son representados por los ceros y unos. Un ejemplo de la codificación y decodificación se encuentra en las Tablas 4.2 y 4.3.

Adicionalmente, para evaluar la homogeneidad de cada columna del alineamiento, calificamos los tres niveles del número enzimático, cada uno como una columna individual (P_{j1}, P_{j2}, P_{j3}, j es el

número de columna en el alineamiento), luego los ponderamos dándole mayor peso a las primeras columnas. La calificación por columna depende de la función objetivo que se haya seleccionado.

	Alineamiento			
	1	**2**	**3**	**4**
S1	1.5.99	6.3.1	1.4.1	1.4.1
S2	1.5.99	-.-.-	-.-.-	1.4.1
S3	1.5.99	1.4.1	1.4.1	6.3.1
S4	1.5.99	6.3.1	1.4.1	1.4.1

Tabla 4.5. Ejemplo de un posible alineamiento

	C1			**C2**			**C3**			**C4**		
S1	1	5	99	6	3	1	1	4	1	1	4	1
S2	1	5	99	-	-	-	-	-	-	1	4	1
S3	1	5	99	1	4	1	1	4	1	6	3	1
S4	1	5	99	6	3	1	1	4	1	1	4	1
Score	F_1P_{11}	F_2P_{12}	F_3P_{13}	F_1P_{21}	F_2P_{22}	F_3P_{23}	F_1P_{31}	F_2P_{32}	F_3P_{33}	F_1P_{41}	F_2P_{42}	F_3P_{43}

Tabla 4.6. Esquema de evaluación de un alineamiento de números enzimáticos.

Los factores que se usan para ponderar los tres niveles son: $F_1 = 15$, $F_2 = 10$ y $F_3 = 1$. Estos factores fueron determinados empíricamente. En las Tablas 4.5 y 4.6 se ilustra el esquema de evaluación.

4.3.4 Función objetivo

La calidad de un alineamiento se basa en tres elementos: el número de *matches* (elementos idénticos alineados), número de *mismatches* (elementos diferentes alineados) y el número de *gaps*. El primer elemento es ponderado positivamente, y los últimos dos sirven para penalizar la calificación al ser multiplicados por factores negativos.

Podemos dividir la puntuación en dos partes, la homogeneidad en las columnas y la penalización de los *gaps*. Mientras más conservadas (más elementos idénticos estén alineados) se encuentren las columnas y los *gaps* sean más concentrados en la matriz del alineamiento, la calidad será mejor. Para cada uno de estos dos elementos existen varios criterios [41].

La suma de pares es uno de los criterios más populares para evaluar la homogeneidad de las columnas, como se mencionó anteriormente. Este es un ejemplo donde los símbolos son los dígitos del 1 al 9 y el símbolo "-" (Tabla 4.7):

***M* =**

	C1	**C2**	**C3**	**C4**	**C5**	**C6**
S1	1	2	4	1	11	-
S2	1	4	4	1	-	-
S3	1	4	-	3	9	-
S4	1	3	4	-	-	1
Score	**6**	**-4**	**-3**	**-7**	**-9**	**-6**

Tabla 4.7. Alineamiento evaluado usando la fórmula de suma de pares. M es la matriz del alineamiento y el score es el esquema que se usa para evaluar los pares de símbolos (ecuación 4.5). Cuando dos *gaps* son evaluados la calificación es cero.

$$score(m_{ij}, m_{kj}) = \begin{cases} 1 \ si \ m_{ij} = m_{kj} \\ -1 \ si \ m_{ij} \neq m_{kj} \\ -2 \ si \ m_{ij} \ ó \ m_{kj} \ es \ un \ gap \end{cases}$$

Ecuación 4.5

En *C4* tenemos:

Score(1,1) + Score(1,3) + Score (1,-) + Score (1,3) + Score (1,-) + Score (3,-) = 1 -1 -2 - 1 -2 -2 = -7

La puntuación (S) del alineamiento considerando solamente la homogeneidad de las columnas seria: *S = 6 -4 -3 -7 -9 -6 = -23.*

Otro criterio utilizado para calificar la homogeneidad de las columnas es el cálculo de la entropía mínima. La entropía de Shannon es una medida de la incertidumbre de un conjunto de datos, depende de la distribución de probabilidad de los mismos. Cuando todos los valores de la variable tienen probabilidades similares la entropía es alta, ya que es difícil poder decir cuál es el siguiente valor de la variable, porque todos son igualmente probables; lo contrario ocurre cuando algunos valores tienen probabilidades mayores que otros [49]. Cuando se usa esta medida como una estrategia para medir la variabilidad en una columna de símbolos alineados, se incorporan las frecuencias y el número de posibles símbolos que existen.

Entonces, la entropía de un alineamiento se puede calcular como la suma de las entropías de las columnas del mismo. La entropía de la columna m_j se define como:

$$E(m_j) = -\sum_a c_j(a) \log_2 p_j(a)$$

Ecuación 4.6

Donde:

m_j = *La j-ésima columna del alineamiento M.*

$c_j(a)$ = *contador del símbolo a en la columna j.*

$p_j(a)$ = *probabilidad del símbolo a en la columna j.*

Para simplificar el problema se considera que todas las secuencias fueron generadas independientemente, y se puede asumir que los símbolos dentro de las columnas y entre ellas, son independientes. La probabilidad $p_j(a)$ puede ser estimada a partir de $c_j(a)$:

$$p_j(a) = \frac{c_j(a)}{\sum_{a'} c_j(a')}$$

Ecuación 4.7

La probabilidad del símbolo *a* en la columna *j* es el cociente del contador del símbolo *a* en la columna *j* entre la suma de los contadores de todos los símbolos (*a'*) que aparecen en la columna *j*. Los *gaps* son considerados como un símbolo más [17].

Entonces, mientras más homogénea sea la columna, menor será el valor de la entropía.

Ejemplo (Tabla 4.8): Tomando el alineamiento *M* presentado anteriormente (Tabla 4.7) tenemos que las entropías serían:

	C1	**C2**	**C3**	**C4**	**C5**	**C6**
Entropía	0	1.80	0.9768	1.80	1.80	0.9768

Tabla 4.8. Evaluación de la entropía

La puntuación del alineamiento tomando en cuenta solo la entropía de las columnas es:

S(M) = 0 + 1.80 + 0.9768 + 1.80 + 1.80 + 0.9768 = 7.35

La penalización de *gaps* contribuye a la evaluación total del alineamiento, usando una calificación negativa. Un *gap* es una secuencia de espacios consecutivos dentro de una sola secuencia en un alineamiento. La longitud del *gap* es el número de *gaps* individuales en él. La relevancia de estos símbolos en el análisis evolutivo de las secuencias se mencionó anteriormente. La idea es tratar el *gap* como un todo, en lugar de dar a cada espacio el mismo peso. Existen varios modelos para la penalización de *gaps*, como son la penalización constante, la afín[8] y la lineal [41].

La penalización *afín* de *gaps* [41], otorga un peso a la apertura de los *gaps* y otro a la extensión. La penalización total para un *gap* de longitud *q* es:

$$P(gap) = W_g + qW_e$$

Ecuación 4.8

W_g = *Penalización por apertura*

q = Longitud del gap.

W_e = *Penalización por extensión*

Otra propuesta, es un criterio que refleja el grado de agrupación de las codificaciones individuales de *gaps* (espacios) en una secuencia en forma de bloques. La justificación biológica es que, desde un punto de vista evolutivo es más parsimonioso encontrar, por ejemplo, un *gap* de cinco posiciones, que cinco *gaps* de un solo espacio cada uno, ya que el primero representa un solo evento. En este criterio de concentración de *gaps* se premia a los alineamientos con una concentración mayor, lo cual implica un número menor de eventos necesarios para explicarlos. Se usa la siguiente fórmula:

$$GC = \frac{\overline{S}_{GB}}{GP}$$

Ecuación 4.9

[8] *affine gap penalty function*

Donde $\overline{S}_{GB}$ es la longitud promedio de un *gap*, y GP es el total de espacios en todo el alineamiento.

Cabe mencionar que los *gaps* iniciales y los *gaps* finales (los que se encuentran al inicio y al final de la codificación de la secuencia) no se toman en cuenta para el cálculo de la penalización. Otra consideración especial se toma cuando no existen *gaps* en el alineamiento (*GP = 0*), en este caso la concentración de *gaps* se hace cero (*GC = 0*) [6].

En este trabajo se consideraron dos diferentes funciones de puntuación para los alineamientos, la primera toma el criterio de minimización de la entropía para evaluar la homogeneidad de las columnas y el criterio de concentración de *gaps*, para la penalización de los mismos; en este caso la función objetivo se debe minimizar. La segunda función utiliza la suma de pares y la penalización *afín* de *gaps*, el objetivo es maximizar la función. Se realizaron ciertas modificaciones a cada uno de estos criterios para adaptarlos al problema particular.

4.3.4.1 Entropía mínima

Esta función está compuesta por tres factores ponderados:

$$Fitness = 0.9*Homogeneidad+0.05*PenalizacionGaps+0.05*IncrementoCols$$

Ecuación 4.10

Los valores devueltos por esta función están entre cero y uno, un valor cercano a cero indica mayor calidad en el alineamiento. Los factores que se utilizaron en la Ecuación 4.10 fueron determinados empiricamente.
Cada uno de estos elementos se describe a continuación.

Homogeneidad

Se calcula la entropía normalizada para cada uno de los niveles de la columna ($E_{Norm1,}$ $E_{Norm2,}$ E_{Norm3}), después se multiplican por los factores correspondientes y se suman. Así, la entropía de la columna se obtiene dividiendo la suma anterior entre la suma de los tres factores. De esta forma se tiene un valor entre 0 y 1, siendo 0 el mejor valor posible y 1 el peor. Ya que los *gaps* son

considerados como un símbolo más en el cálculo de la entropía, también se calcula una penalización por el número de *gaps* presentes en la columna. Esta penalización extra fue necesaria debido a que es posible que la mayoría de símbolos en una columna sean *gaps* y se tenga una buena evaluación, ya que es homogénea. El número de *gaps* se divide entre el número máximo de *gaps* posibles en una columna (total de secuencias menos uno), para obtener un valor entre 0 (sin *gaps*) y 1 (máximo número de *gaps*). Estos dos valores (entropía y penalización de *gaps* por columna) son ponderados para obtener así el score final de la columna

$$Homogeneidad_i = E_i * 0.6 + GapsCol_i * 0.4$$

Ecuación 4.11

Donde:

$$Homogeneidad_i = Score\ de\ la\ homogeneidad\ en\ la\ columna\ i$$

$$E_i = \frac{E_{Normi1} * F_1 + E_{Normi2} * F_2 + E_{Normi3} * F_3}{F_1 + F_2 + F_3}$$

$$GapsCol_i = \frac{\#Gaps}{\#Sec - 1}$$

Así la calificación del alineamiento, tomando en cuenta solo la homogeneidad, está dada por el promedio de las puntuaciones de todas las columnas.

Penalización de *gaps*

El criterio de concentración de *gaps* fue elegido para esta penalización (Ecuación 4.9). Para este cálculo contamos todos los *gaps* (bloques continuos de *gaps* individuales) y los tamaños de los mismos, para cada secuencia del alineamiento. Después obtenemos el promedio del tamaño de estos *gaps* y dividimos entre el número total de *gaps* individuales, como se indica en la fórmula mencionada anteriormente. Finalmente obtenemos un valor entre 0 y 1, donde un valor cercano a cero indica que los *gaps* están concentrados en un número menor de bloques largos, y un valor cercano a uno indica que los *gaps* se encuentran en varios bloques de longitudes cortas.

Incremento de columnas

Este factor penaliza el aumento de columnas en la matriz del alineamiento, que es un resultado de la inserción de *gaps*. Se calcula con la siguiente fórmula [6]:

$$IncrementoCols = \frac{MAX}{CAlig}$$

Ecuación 4.12

Donde:

MAX = La longitud de la secuencia más larga del alineamiento sin tomar en cuenta los gaps.

CAlig = El número de columnas del resultado final del alineamiento.

Este valor se encuentra entre 0 y 1; mientras más cercano a 1 el alineamiento recibirá una penalización mayor.

4.3.4.2 Suma de pares

Esta función está determinada por tres factores:

$$Fitness = Homogeneidad + PenalizaciónGaps - IncrementoCols$$

Ecuación 4.13

Los factores de la fórmula no son ponderados como en la función de la entropía mínima. En este caso se trata de maximizar el valor de la función. Los valores no se encuentran normalizados como en el caso anterior. Los tres factores se describen a continuación.

Homogeneidad

Este valor se calcula con la fórmula de suma de pares mencionada anteriormente (Ecuación 3.1). En este caso los factores de puntuación que se utilizaron son:

$$s(m_{ik}, m_{jk}) = \begin{cases} 2 \ si \ los \ símbolos \ son \ iguales \\ -1 \ si \ los \ símbolos \ son \ diferentes \\ -2 \ si \ uno \ de \ los \ símbolos \ es \ un \ gap \\ 0 \ si \ los \ dos \ símbolos \ son \ gaps \end{cases}$$

Ecuación 4.14

Donde:

$s(m_{ik},m_{jk})$= Es la puntuación de los símbolos en las filas *i* y *j*, para la columna *k.*

De igual forma que en el cálculo de la entropía, la suma de pares se realiza para cada uno de los tres niveles de cada columna. La homogeneidad de la columna es el resultado de la suma de las puntuaciones obtenidas por la suma de pares, de cada uno de los niveles. La calificación del alineamiento es la suma de la puntuación de cada columna.

Penalización de *gaps*

En este caso se utilizó el criterio de penalización *afín* de *gaps*, los factores fueron:

AperturaGap = -2

ExtensiónGap = -0.125

El valor se calcula usando la fórmula descrita anteriormente (Ecuación 4.8).

Incremento de columnas

Se utiliza el mismo criterio empleado en el caso de la entropía (Ecuación 4.12), solamente que se pondera por un factor igual a 2.

Los valores de los parámetros utilizados en los tres factores fueron determinados empíricamente.

4.4 Algoritmo progresivo

Dentro de los enfoques existentes para el alineamiento múltiple de secuencias biológicas uno de los más populares es el alineamiento progresivo. Este método comienza con el alineamiento de las dos secuencias que están más cercanas evolutivamente (esto se determina con un análisis previo de alineamiento por pares) y sucesivamente se agrega la siguiente secuencia más cercana al par inicial. Este proceso continua iterativamente, ajustando los *gaps* dentro de todas las secuencias. *MUSCLE* [19] es una de las herramientas más populares para el alineamiento de secuencias que utiliza este enfoque, con muy buenos resultados [8].

Este enfoque se implementó para realizar los alineamientos de vías metabólicas, basándonos en el algoritmo genético que se ha descrito. El primer paso consistió en realizar los alineamientos por pares de las 452 secuencias de la base de datos, usando el algoritmo genético. Con estos resultados se crea una matriz de distancia (valores de similitud de secuencia), para tener la relación de cada secuencia con las demás. Estos valores son los que nos sirven para guiar el alineamiento múltiple,

empezando con las secuencias más cercanas entre sí. En cada iteración se agrega una secuencia al alineamiento anterior, y en este momento es donde el algoritmo genético interviene para mejorar el alineamiento.
Entonces tenemos una llamada al algoritmo genético por cada secuencia que se alinea, y el algoritmo genético a su vez es ejecutado por un número de generaciones definido dentro de los parámetros del algoritmo.

4.5 Análisis evolutivo de las secuencias

Los alineamientos múltiples se realizan con el objetivo de encontrar relaciones evolutivas en secuencias que en un principio no tienen indicios de relación alguna. Dentro de la metodología de este trabajo, se propone primero crear grupos de secuencias "parecidas" para posteriormente (usando el algoritmo para el alineamiento múltiple) poder identificar específicamente cuáles son los segmentos de las secuencias que son similares entre ellas. Para esta tarea se utilizó una técnica de aprendizaje no supervisado descrita a continuación.

4.5.1 Clustering

El principal objetivo de las técnicas de aprendizaje no supervisado es encontrar agrupamientos naturales o particiones que sean significativas usando una función de distancia. El *clustering* o agrupamiento es una técnica que ha sido ampliamente usada en gran variedad de campos, incluida la biología. En el análisis de secuencias, el *clustering* es usado para agrupar secuencias homólogas en familias de proteínas [20].

4.5.1.1 K-Means

Existen varios algoritmos de *clustering* basados en técnicas de partición. Uno de los más populares es el algoritmo *K-Means*, debido a que es muy fácil de implementar en muchos problemas prácticos. El algoritmo es el siguiente:
Considere el conjunto *S* con *n* objetos:

$$S = \{x_i : 1 \leq i \leq n\}$$

1) Inicializar una *partición-k* aleatoriamente o basada en alguna información, eligiendo *k* centros.

$\{C_1, C_2, \ldots, C_k\}$

2) Asignar cada elemento del conjunto al *cluster* más cercano (al centro más cercano).

$$x_j \in C_m \ si \ \left\| x_j - C_m \right\| \leq \left\| x_j - C_i \right\|$$

$$\forall j,\ 1 \leq j \leq k,\ j \neq m, donde \ \ j = 1,2,\ldots,n$$

3) Recalcular los nuevos centros de los grupos.
4) Repetir 2) y 3) hasta que no haya cambios en los *clusters*.

Este algoritmo es bastante eficiente, pero tiene como desventaja que el número de *clusters k* se debe establecer a priori [29].

Determinar la *k* óptima es un problema difícil y no existe una regla para la mejor elección. Existen varios criterios que se pueden usar, uno de los más comunes es el llamado *"criterio del codo"*. Consiste en graficar una "medida de calidad" del *clustering* contra el número de *clusters* y buscar un "codo" en esta curva. Se examina cómo cambia la calidad del *clustering* en función del número de *clusters k.* Este criterio selecciona la *k,* tal que al aumentar el número de *clusters* no agrega suficiente información. Entonces buscamos la *k* donde ocurra el cambio más significativo, para lo cual seleccionamos la *k* que maximiza la segunda derivada de la función [51,56].

El algoritmo *k-means* se implementó para crear *clusters* de secuencias que fueran parecidas entre ellas, usando como medida de distancia la similitud entre las secuencias, calculada al realizar los alineamientos por pares. Para asignar los elementos se calcula el promedio de similitud con todos los elementos del *cluster*, y se asigna al grupo con el que se obtenga el menor promedio. Los centros son recalculados obteniendo la similitud promedio de cada elemento del *cluster* con el resto de los elementos, la secuencia que tenga el menor promedio de similitud, será el nuevo centro del *cluster*.

Para evaluar la calidad del *clustering*, al final de la iteración se evalúa la calidad de cada grupo, calculando la similitud de los centros con todos los elementos del *cluster*. Conforme se va iterando, este valor disminuye, ya que los grupos tendrán elementos más parecidos entre sí. Esta medida de calidad nos sirve para elegir el número de *clusters* óptimo, siguiendo el "criterio del codo". Se graficó esta medida contra varios valores de *k,* que van de dos a 100, es decir, se realizó el *clustering* de las 452 secuencias probando desde dos grupos hasta 100. Los experimentos realizados para elegir la *k* óptima se describen en el siguiente capítulo.

Los grupos obtenidos con el *clustering* fueron posteriormente depurados para obtener una mayor calidad. Las secuencias que tuvieran una similitud promedio, con el resto de los elementos del grupo, mayor a 0.4 fueron eliminadas del *cluster*. Estas secuencias no tenían una similitud significativa con ninguna otra.

4.6 Visualizador

Después de realizar los alineamientos múltiples sobre estos grupos de secuencias, el resultado es presentado en una herramienta que permite visualizar las enzimas alineadas (números enzimáticos), usando seis rangos de colores para diferenciar las seis categorías del primer nivel de los números enzimáticos. Por ejemplo, los números enzimáticos cuyo primer nivel sea el 1, están representados por varias tonalidades del color rojo, el 1.1.1 siendo el tono más débil y de ahí se va aumentado la tonalidad dependiendo del número total de enzimas que tengan el 1 en el primer nivel. De esta manera es fácil identificar cuáles son las enzimas similares dentro del alineamiento (Figura 4.4).

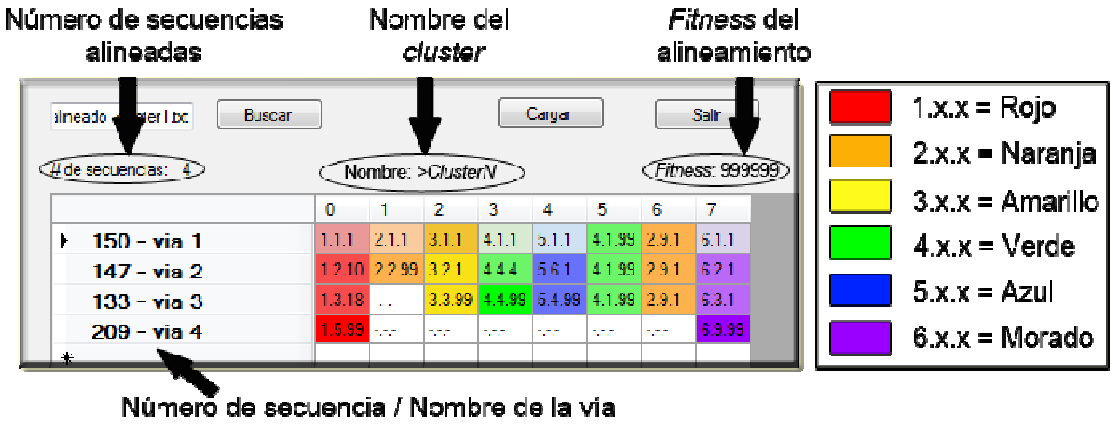


Figura 4.4. Las columnas 5 y 6 tienen elementos idénticos por lo cual aparecen del mismo color. La primera columna indica el identificador de la secuencia (del 1 al 452) y el nombre de la vía a la que pertenece. También se muestra el número del *cluster* y el *fitness* del alineamiento.

Capítulo 5

RESULTADOS

5.1 Vías metabólicas y secuencias

Las 452 secuencias obtenidas a partir de la reconstrucción de las vías metabólicas utilizando el algoritmo BFS se muestran en el apéndice B. En la Figura 5.1 se muestra una gráfica de la distribución de estas secuencias y de las longitudes promedio (pasos enzimáticos) de cada una de las secuencias obtenidas. De estas gráficas podemos deducir que el número promedio de secuencias por ruta metabólica es de 11, siendo el metabolismo del piruvato la ruta que presenta la mayor cantidad de secuencias (145). La ruta 13, el metabolismo de la pirimidina, presenta alrededor de 40 secuencias diferentes, siendo la segunda más abundante dentro de todas las secuencias obtenidas.

Adicionalmente, se identificó que la longitud promedio de las secuencias es de 6.22 pasos enzimáticos, siendo la ruta 45 (metabolismo del piruvato) la que presenta las secuencias más largas de todo el conjunto. También se observó la existencia de 11 vías con una sola secuencia y 5 que solamente tienen dos (Figura 5.1b).

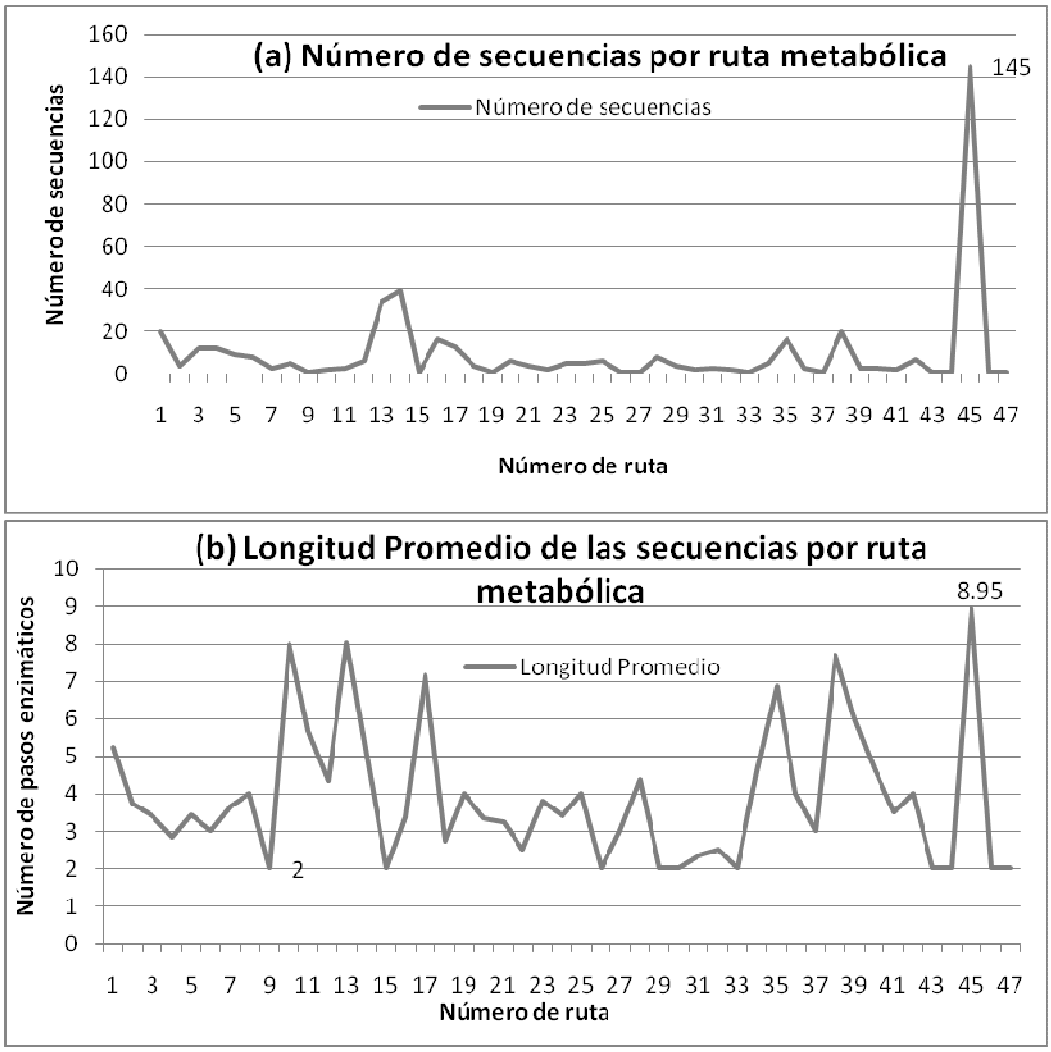


Figura 5.1. El apéndice A presenta la relación de las rutas metabólicas y sus identificadores. En el panel (a) se presenta el número de secuencias obtenidas por cada mapa. En el panel (b) se muestra la longitud promedio de las secuencias. El metabolismo del piruvato es la ruta que presenta la mayor cantidad de secuencias.

5.2 Alineamientos por pares

Los primeros experimentos que se realizaron fueron los alineamientos por pares que posteriormente formaron parte del algoritmo de *clustering* y de los alineamientos múltiples. El objetivo de dichos alineamientos es la generación de grupos de secuencias con segmentos en común, pero que no son evidentes a priori.

Las 452 secuencias de las 47 rutas metabólicas de *E. coli* fueron comparadas entre sí; el alineamiento por pares se realizó usando el AG descrito anteriormente, con la función objetivo que evalúa la entropía de las columnas. Los parámetros fueron los siguientes:
Tamaño de población = 100 individuos
Tasa de mutación = 0.01
Tasa de cruza = 0.90

El algoritmo se detenía cuando se hubieran ejecutado 20 generaciones sin cambio en el valor de la función objetivo. Se realizaron 10 repeticiones del AG y se tomó el mejor resultado. En el apéndice C se muestra la matriz de todos los alineamientos. La Figura 5.2 muestra un segmento de esta matriz. En el eje horizontal se muestran las secuencias de la 200 a la 315 y en el eje vertical de la 235 a la 345. La barra de colores indica el nivel de similitud del par de secuencias alineado, mientras más cercano a cero (azul) el alineamiento es mejor, es decir, las secuencias comparten una mayor cantidad de elementos conservados.

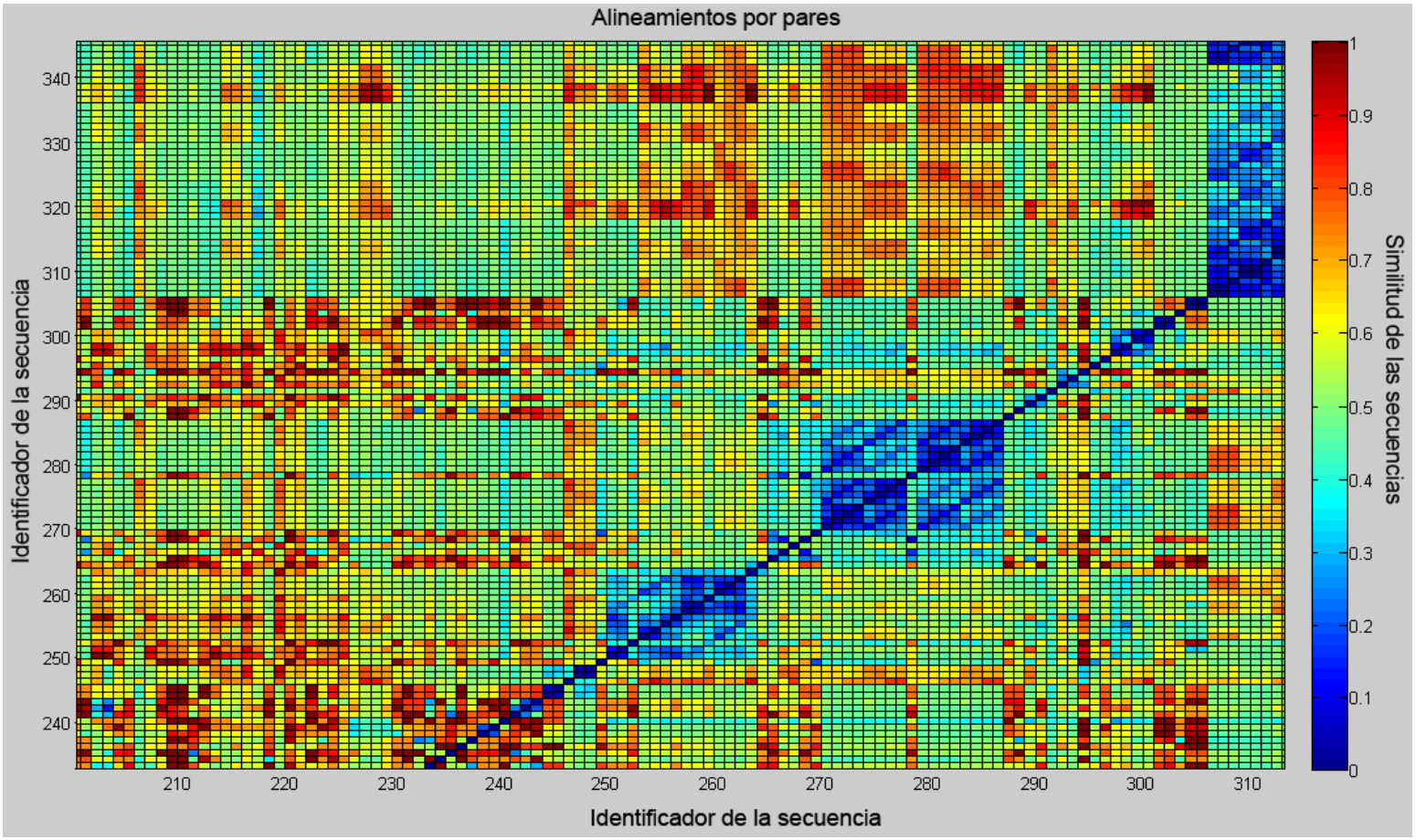


Figura 5.2. Los ejes horizontal y vertical tienen los identificadores de las secuencias. La barra de la derecha indica el nivel de similitud del alineamiento por pares (de 0 a 1, donde 0 es el mejor valor). Se observa alrededor de la diagonal principal la concentración de recuadros azules, indicando la similitud de las secuencias dentro de una misma vía, y también podemos ver que existen dispersos en toda la matriz, recuadros indicando una similitud alrededor de 0.4.

5.3 Clustering

Para seleccionar el número de *clusters* (el parámetro *k* en el algoritmo *K-Means*), se utilizó el criterio del codo, graficando la calidad del *cluster* para *k = 2* hasta *100*. Se realizaron 20 repeticiones para cada *k* y se seleccionó el mejor resultado (Figura 5.3).

Encontrar el codo en la Figura 5.3 es difícil, por lo que se empleó la gráfica de la segunda derivada de esta función (Figura 5.4 (a) y (b)), identificando la *k* que maximiza este valor, ya que representa la mayor pendiente (el cambio más grande entre dos *k's* diferentes). La Fígura 5.4a presenta la función con mucho ruido debido a la cantidad de puntos que se grafican, por lo que en la Fígura 5.4b se presenta el intervalo donde se tienen los picos mas importantes y se interpolaron los valores para obtener una gráfica mas suave. Los valores de la función se escalaron para observar mejor los cambios presentes.

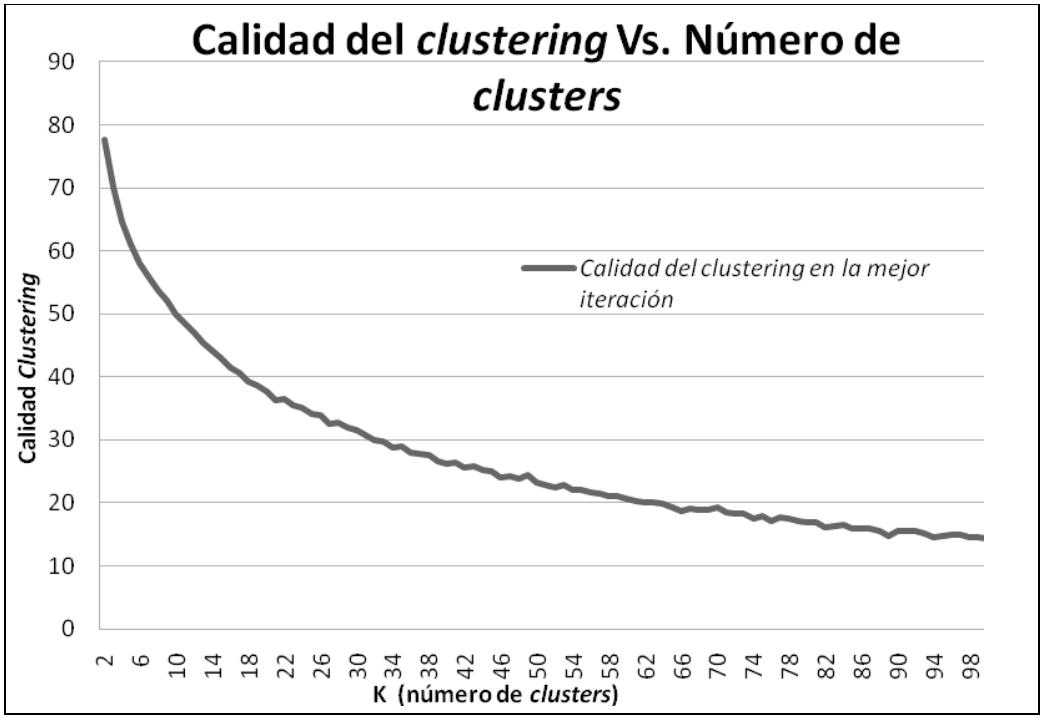


Figura 5.3. Calidad del *clustering* para *k*=2 hasta 100. Se puede observar como al aumentar el número de *clusters* la calidad se incrementa (se está minimizando), ya que los *clusters* contienen cada vez menos secuencias diferentes.

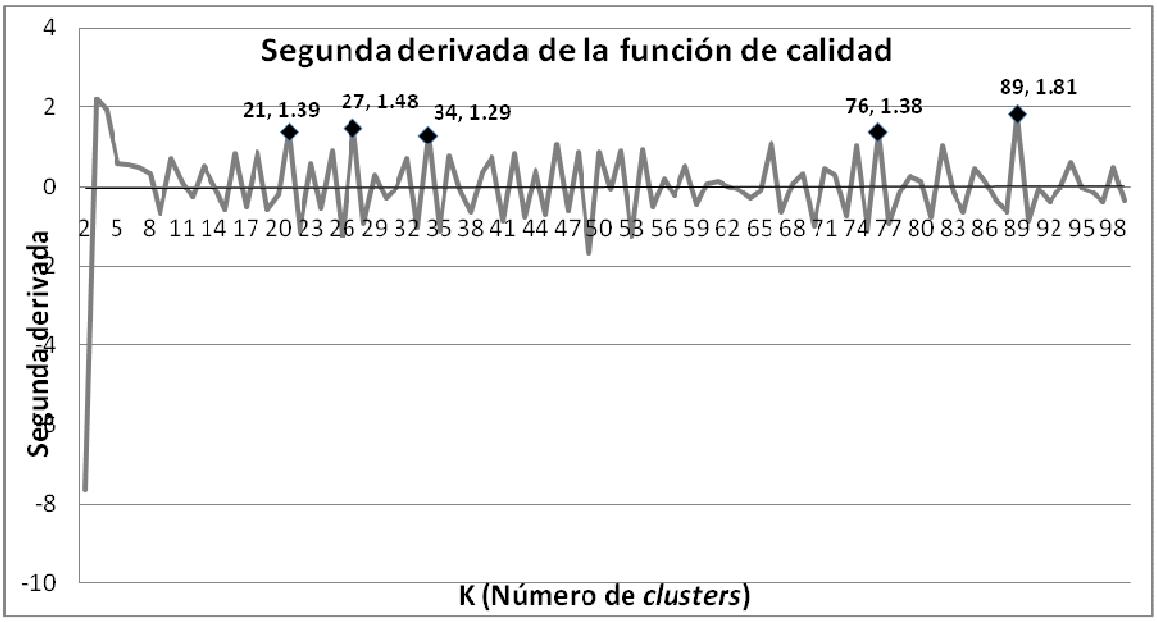


Figura 5.4 (a). Gráfica de la segunda derivada de la función de calidad

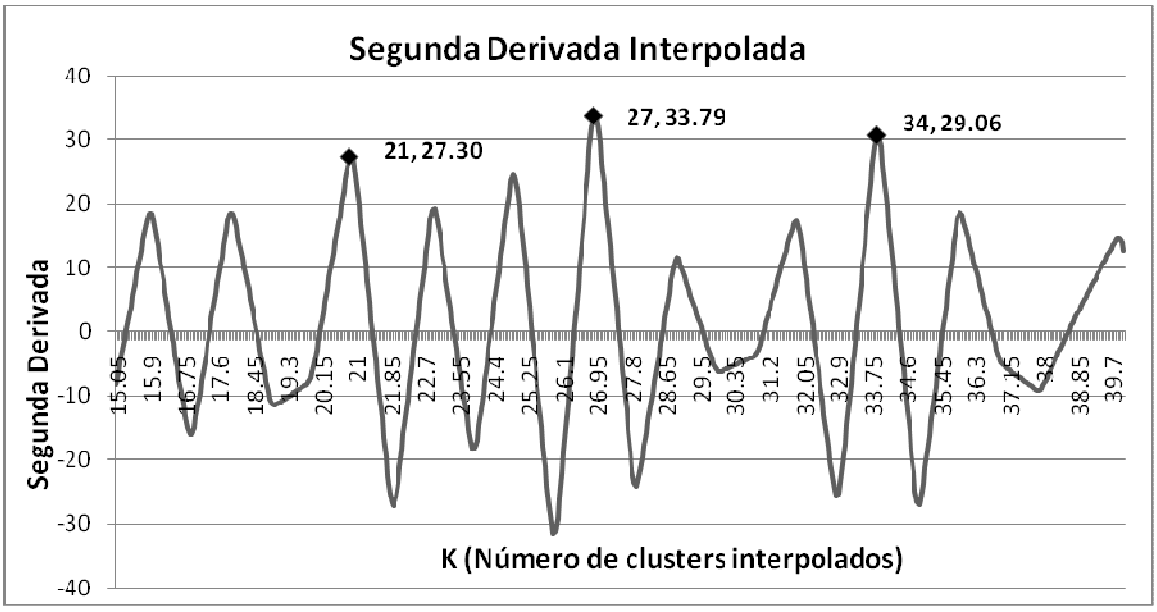


Figura 5.4 (b). Gráfica interpolada de la segunda derivada de la función de calidad. Se grafican solamente los valores de k del 15 al 40.

El cambio más notable en la calidad del *clustering* se muestra en *k=3*, ya que en el punto anterior el agrupamiento se generó aleatoriamente, y en el siguiente se empieza a usar la información descubierta por el algoritmo. Pero este punto no se puede tomar ya que solamente se tendrían tres grupos y la calidad no es la óptima. Dado lo anterior, diversos

valores de *k* se pueden elegir para generar el *clustering*; los siguientes fueron elegidos: 21, 27 y 34 (se pueden observar mas claramente en la Figura 5.4b), ya que presentan picos altos en la gráfica. Los puntos 76 y 89 también presentaron valores altos, pero fueron descartados ya que se sabe que al acercarse el número de *clusters* al número de elementos totales la calidad es mayor, debido a que los grupos cada vez contienen menos secuencias hasta llegar a grupos con solamente una o dos secuencias.

Se realizaron pruebas con el *clustering* de 21, 27 y 34 grupos, finalmente se eligió *k=34* ya que para los primeros dos valores (21 y 27), al realizar el proceso de depuración se encontraron 93 y 71 secuencias respectivamente, que tenían un valor de similitud mayor al umbral ≥ 0.4. Para *k = 21* estas secuencias aparecían en 19 de los 21 *clusters*, y para *k = 27* en 21 de los 27 *clusters*. Al elegir *k = 34*, este número se redujo a 37 secuencias distribuidas en 18 *clusters*, con lo cual se obtuvo una mejor calidad en el proceso de agrupamiento.

La distribución de las vías dentro de los *clusters* se encuentra en el apéndice B. La Figura 5.5 muestra un histograma con la distribución de los *clusters* antes de ser depurados. Los colores representan las 47 diferentes vías, cada barra es un grupo. El tamaño de cada barra representa el número de secuencias que integran cada grupo, mientras que los colores indican las diferentes vías que están incluidas en cada *cluster*. De esta gráfica se observa que el *cluster* 3 es el más grande (24 secuencias fueron incluidas), mientras que el grupo 16 es el más pequeño con solamente 3 secuencias.

De manera alternativa, los grupos 2 y 30 son los grupos que contienen el mayor número de vías diferentes (9 vías), por lo que fueron considerados como los más heterogéneos.

En cuanto a los resultados biológicos encontrados se pueden hacer varias observaciones:

- El piruvato es uno de los productos finales donde convergen al menos siete diferentes rutas metabólicas, lo cual podría explicar su presencia en nueve *clusters* diferentes. Es decir, es una de las rutas más antiguas en el metabolismo celular y su parecido a diversas rutas alternas podría sugerir que las otras vías han sido originadas a partir del piruvato.
- Las secuencias que aparecen en varios *clusters* sugieren que se han reclutado a partir de otras vías metabólicas pre-existentes (más antiguas). Es decir, la vía del

piruvato estaría conformada por fragmentos de alrededor de 9 rutas diferentes y/o que pueden estar compartidas con otras vías metabólicas. Sin embargo, existen fragmentos que se han originado casi exclusivamente para la síntesis de piruvato (*clusters* 7, 8, 27 y 34).

- Existen casos donde las secuencias se presentan en un solo *cluster*. Por ejemplo, el *cluster* 26 contiene solo dos tipos de secuencias asociadas a dos rutas metabólicas: el metabolismo del glutatión (*Glutathione metabolism*) y la biosíntesis del peptidoglicano (*Peptidoglycan biosynthesis*). Una hipótesis para explicar estos casos es que ambos metabolismos compartieron o comparten una vía común y a partir de algún punto de la secuencia, divergen.

- Alternativamente, hay *clusters* donde existe una amplia diversidad de secuencias. Por ejemplo, el *cluster* 2 contiene nueve diferentes vías metabólicas (15 secuencias), sin embargo la longitud promedio de las secuencias es pequeña (3.45). A partir de estos datos, se observa la siguiente tendencia: a menor número de vías asociadas a los *clusters*, mayor será la longitud promedio de la secuencia (ver apéndice D). Este dato sugiere que a menor tamaño de la secuencia mayor "promiscuidad", es decir, las vías metabólicas están reclutando preferencialmente secuencias cortas para incrementar su tamaño o bien, desde un punto de vista funcional, convergen o son el punto de partida hacia un producto en particular.

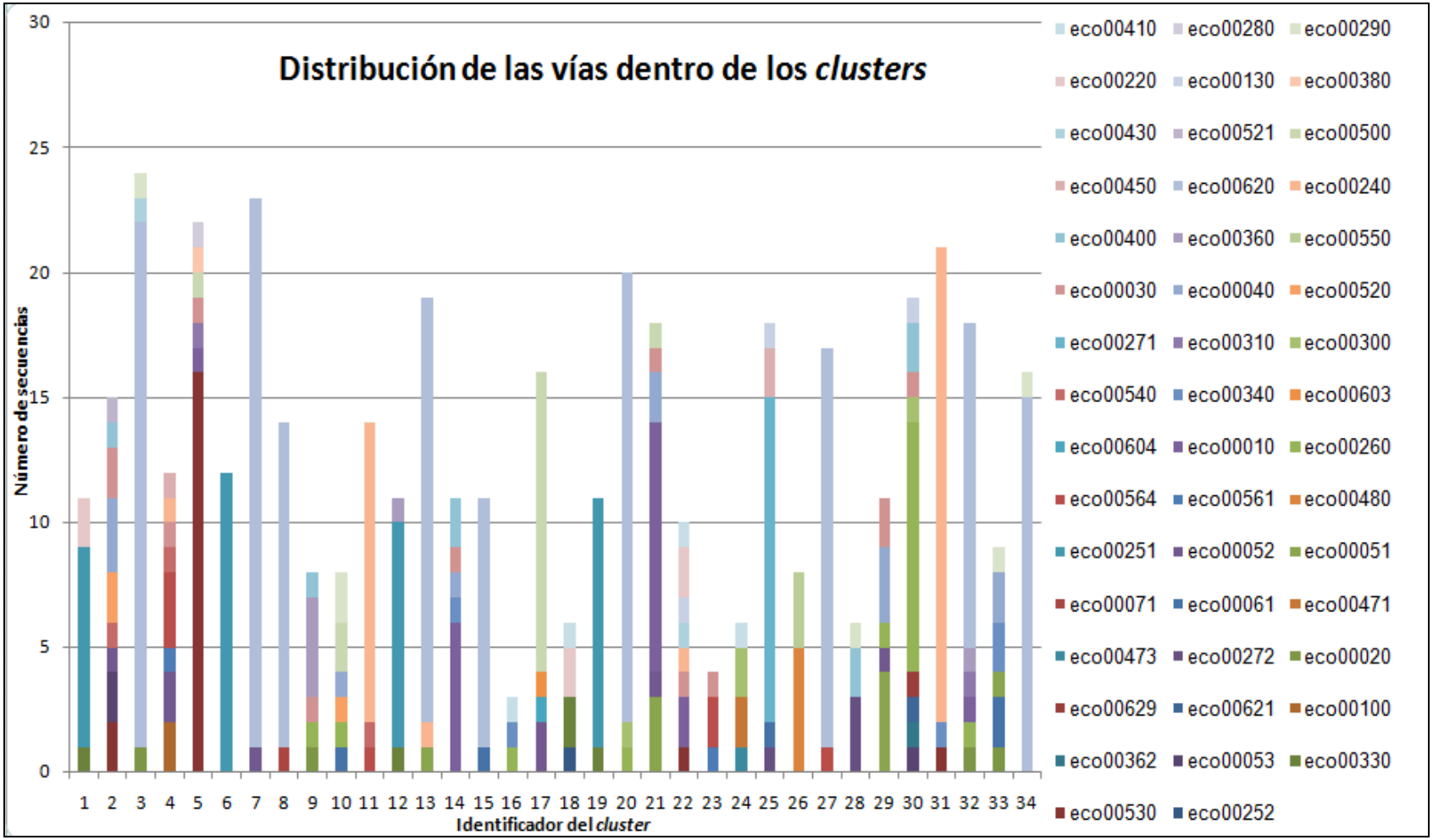


Figura 5.5. Distribución de las 452 secuencias en los 34 *clusters*.

5.4 Alineamientos múltiples

Para los alineamientos múltiples se realizaron cuatro experimentos probando las dos funciones objetivos con dos versiones del AG, el normal y el AG combinado con el alineamiento progresivo. Los mejores resultados se obtuvieron usando el *"AG progresivo"* y la función objetivo basada en la entropía, aunque en algunos casos no hubo diferencia significativa entre las dos funciones objetivos.

Para la suma de pares es necesario definir los valores con los que se evaluarán los alineamientos de símbolos iguales, símbolos diferentes y con los *gaps*. En este sentido, se realizaron diversas pruebas, observando cambios significativos al modificar estos valores, finalmente se usaron los valores que se presentaron en el capítulo anterior. La función basada en la entropía usa tres factores para ponderar el nivel del número enzimático al calcular la entropía (15, 10 y 1), se probaron varios valores para estos factores y no se encontraron cambios significativos, incluso eliminando los factores. Cabe mencionar que la función objetivo es un parámetro de entrada del programa desarrollado, y es posible agregar posteriormente otras diferentes a las propuestas en este trabajo.

En el caso de los dos AG's probados se encontraron diferencias significativas en los alineamientos, con las dos funciones objetivo que se proponen. El *AG normal* obtenía alineamientos con muchos *gaps* individuales, que ocasionaban una alta calificación (la función se debe minimizar) en la evaluación. Se evaluaron los distintos operadores de mutación para tratar de evitar este problema, así como distintos factores para la penalización de la inserción de *gaps*, pero no se logró mejorar significativamente.

El *"AG progresivo"* logró solucionar este problema, primero ordenando las secuencias según su similitud con las demás, usando este orden como guía para realizar el alineamiento de las secuencias empezando por las dos más similares y agregando una a la vez. Este algoritmo resultó ser más lento que el primero, pero los resultados son mucho más significativos como se puede observar en la Figura 5.6, donde se compara el *fitness* de los alineamientos de los 34 *clusters* utilizando ambos algoritmos con la función objetivo basada en la entropía.

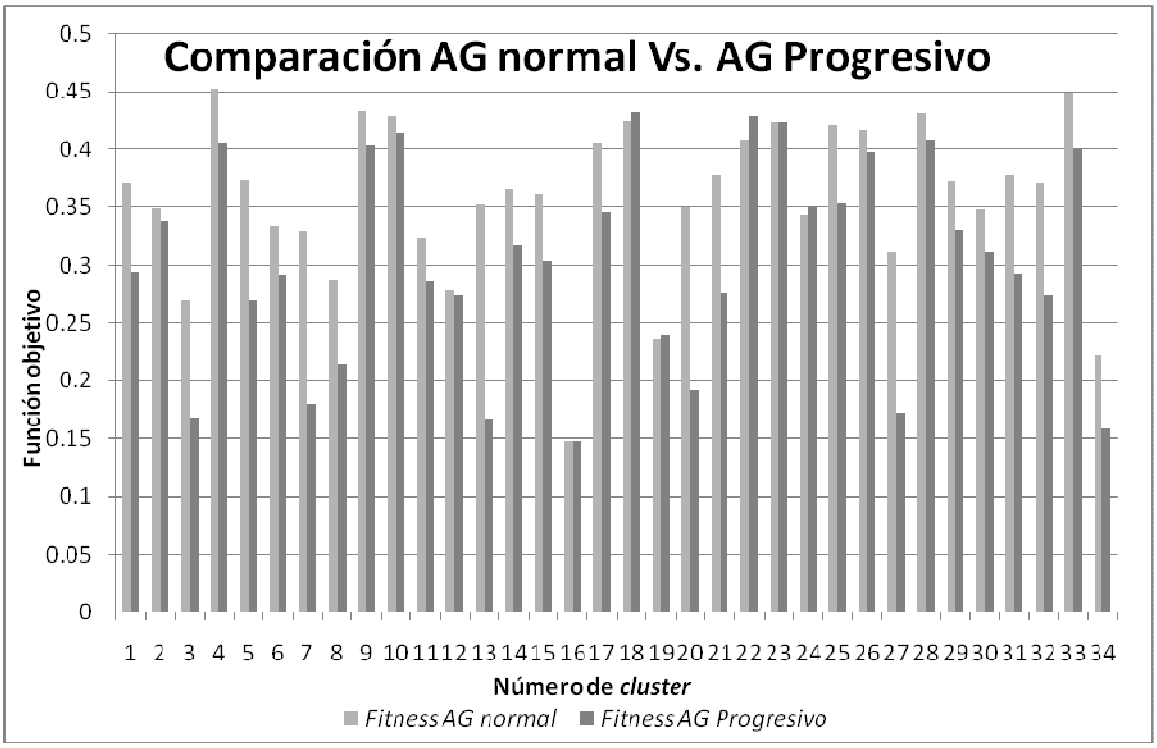


Figura 5.6. En prácticamente los 34 alineamientos, el AG progresivo obtuvo mejores resultados que el AG normal y en los casos donde no fue así, los resultados fueron muy cercanos.

Otra diferencia entre los resultados de los dos AG's se encuentra en la minimización de eventos que se representan con los *gaps*, recordemos que se prefiere tener *gaps* concentrados a *gaps* individuales, ya que es más factible la ocurrencia de un solo evento evolutivo explicado con un *gap* de tamaño 4 a la ocurrencia de 4 eventos distintos, usando 4 *gaps* individuales. El alineamiento del *cluster* 13 (Figuras 5.7 y 5.8) ilustra los resultados mencionados anteriormente. Los alineamientos obtenidos con el AG normal (con las dos funciones objetivo) tienen entre 6 y 9 columnas más que los obtenidos con el AG progresivo. Se puede observar también el gran número de *gaps* individuales dispersos a lo largo del alineamiento (Figura 5.7b). Los resultados del AG progresivo son muy similares entre sí para las dos funciones (Figura 5.8), y existe una diferencia significativa con los presentados en la Figura 5.7. La guía que se utilizó para realizar estos alineamientos logró

minimizar el número de *gaps* insertados (representación de los eventos evolutivos) y alinear el mayor número de columnas idénticas.

La evaluación de los alineamientos con los cuatro experimentos se encuentra en la Tabla 5.1. Para las dos funciones objetivo, el AG progresivo obtuvo los mejores resultados. El resultado del AG normal usando la suma de pares fue un valor negativo, lo que refleja el gran número de *gaps* individuales que se insertaron, ya que el alineamiento de un símbolo con un *gap* es calificado negativamente.

AG / Función objetivo	Entropía	Suma de pares
Normal	0.3517	-6474.32
Progresivo	0.1672	5471

Tabla 5.1. Resultados de la evaluación de los alineamientos del *cluster* 13.

(a) Entropía

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
327 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
328 - Pyruvate metabolism	-.-.-	-.-.-	1.1.1	-.-.-	-.-.-	-.-.-	1.2.1	1.1.2	-.-.-	-.-.-	2.7.9	-.-.-	4.1.1	4.1.1	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
329 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
435 - Pyruvate metabolism	1.1.1	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
436 - Pyruvate metabolism	1.1.1	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
364 - Pyruvate metabolism	3.1.2	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
309 - Pyruvate metabolism	4.2.3	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
310 - Pyruvate metabolism	4.2.3	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
345 - Pyruvate metabolism	4.4.1	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
437 - Pyruvate metabolism	1.1.1	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
346 - Pyruvate metabolism	4.4.1	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	-.-.-	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
365 - Pyruvate metabolism	3.1.2	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	-.-.-	1.2.2	2.7.2	3.6.1	-.-.-	6.2.1	-.-.-	6.2.1	6.4.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
311 - Pyruvate metabolism	4.2.3	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
381 - Pyruvate metabolism	2.3.1	2.7.9	-.-.-	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
382 - Pyruvate metabolism	2.3.1	-.-.-	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
347 - Pyruvate metabolism	4.4.1	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
383 - Pyruvate metabolism	2.3.1	2.7.9	-.-.-	4.1.1	4.1.1	2.7.1	-.-.-	1.2.2	2.7.2	3.6.1	-.-.-	6.2.1	-.-.-	6.2.1	6.4.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
053 - Fructose and mannose metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.7	9.9.9	-.-.-
119 - Pyrimidine metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	6.3.5	9.9.9

(b) Suma de pares

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
327 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	-.-.-	6.2.1	6.2.1	2.3.1	-.-.-	-.-.-	-.-.-	-.-.-
328 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	-.-.-	4.1.1	2.7.1	1.2.2	-.-.-	2.7.2	-.-.-	3.6.1	-.-.-	6.2.1	-.-.-	-.-.-	6.2.1	2.3.3
329 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	-.-.-	4.1.1	4.1.1	2.7.1	1.2.2	-.-.-	2.7.2	3.6.1	6.2.1	6.2.1	-.-.-	-.-.-	6.4.1	-.-.-	-.-.-
435 - Pyruvate metabolism	-.-.-	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	-.-.-	2.7.1	-.-.-	1.2.2	-.-.-	2.7.2	3.6.1	6.2.1	-.-.-	6.2.1	2.3.1	-.-.-
436 - Pyruvate metabolism	1.1.1	-.-.-	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	-.-.-	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
364 - Pyruvate metabolism	-.-.-	3.1.2	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	-.-.-	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
309 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	-.-.-	6.2.1	6.2.1	2.3.1	-.-.-	-.-.-	-.-.-	-.-.-
310 - Pyruvate metabolism	-.-.-	4.2.3	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	-.-.-	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-
345 - Pyruvate metabolism	-.-.-	4.4.1	1.2.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	-.-.-	-.-.-	3.6.1	6.2.1	6.2.1	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1
437 - Pyruvate metabolism	1.1.1	-.-.-	1.2.1	2.7.9	4.1.1	4.1.1	-.-.-	2.7.1	1.2.2	2.7.2	-.-.-	-.-.-	3.6.1	-.-.-	6.2.1	6.2.1	6.4.1	-.-.-	-.-.-
346 - Pyruvate metabolism	-.-.-	4.4.1	1.2.1	2.7.9	-.-.-	-.-.-	4.1.1	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	2.7.2	3.6.1	6.2.1	6.2.1	-.-.-	2.3.3	-.-.-
365 - Pyruvate metabolism	3.1.2	-.-.-	-.-.-	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	-.-.-	6.4.1	-.-.-	-.-.-
311 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	-.-.-	-.-.-	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	-.-.-	6.2.1	-.-.-	-.-.-	6.4.1
381 - Pyruvate metabolism	-.-.-	2.3.1	2.7.9	-.-.-	4.1.1	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	-.-.-	-.-.-	2.7.2	3.6.1	6.2.1	6.2.1	-.-.-	-.-.-	2.3.1
382 - Pyruvate metabolism	2.3.1	-.-.-	2.7.9	4.1.1	-.-.-	-.-.-	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
347 - Pyruvate metabolism	4.4.1	1.2.1	2.7.9	-.-.-	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	-.-.-	-.-.-	6.2.1	6.2.1	-.-.-	-.-.-	6.4.1	-.-.-	-.-.-
383 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	-.-.-	-.-.-	1.2.2	-.-.-	2.7.2	3.6.1	6.2.1	-.-.-	-.-.-	6.2.1	6.4.1	-.-.-
053 - Fructose and mannose metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.7	-.-.-	-.-.-	-.-.-	9.9.9	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
119 - Pyrimidine metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	6.3.5	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	9.9.9	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-

Figura 5.7. Alineamiento del *cluster* 13 obtenido con el AG normal.
(a) Función objetivo basada en la entropía.
(b) Función objetivo basada en la suma de pares.
En ambos alineamientos se observa el aumento en el número de columnas debido a la inserción de *gaps* individuales. En el panel (a) se tiene el "mejor" resultado basado en la identidad de las columnas, pero no en el aumento de columnas.

(a) Entropía

	0	1	2	3	4	5	6	7	8	9	10	11	12
327 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
328 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
329 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
435 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
436 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
364 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
309 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
310 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
345 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
437 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
346 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
365 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
311 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
381 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
382 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
347 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
383 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
053 - Fructose and mannose metabolism	-.-.-	-.-.-	-.-.-	2.7.7	-.-.-	-.-.-	-.-.-	9.9.9	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
119 - Pyrimidine metabolism	-.-.-	-.-.-	-.-.-	6.3.5	-.-.-	-.-.-	-.-.-	9.9.9	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-

(b) Suma de pares

	0	1	2	3	4	5	6	7	8	9	10	11	12
327 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
328 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
329 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
435 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
436 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
364 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
309 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
310 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
345 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
437 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
346 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
365 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
311 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
381 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
382 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
347 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
383 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
053 - Fructose and mannose metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.7	9.9.9	-.-.-	-.-.-	-.-.-
119 - Pyrimidine metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	6.3.5	9.9.9	-.-.-

Figura 5.8. Alineamientos del *cluster* 13 obtenidos con el AG progresivo.
(a) Función basada en la entropía
(b) Función basada en la suma de pares
Las diferencias con los alineamientos del AG normal son evidentes. Se logran mejores alineamientos con el AG progresivo para las dos funciones objetivo, y casi no hay diferencias entre estos resultados, únicamente en las últimas dos secuencias. Después del proceso de depuración de los *clusters* estas dos secuencias fueron eliminadas, obteniendo un alineamiento de mayor calidad.

5.5 Mapeo de los alineamientos a las vías metabólicas.

Los alineamientos múltiples nos permitieron identificar módulos (secuencias de reacciones enzimáticas) que se encuentran repetidos en distintas vías metabólicas. Es importante también encontrar en qué parte de la vía se ubican estos módulos, ya que nos podría indicar la existencia de vías alternas para la síntesis de algún producto específico, en el caso de encontrarlas al final o al inicio.

	0	1	2	3	4	5	6	7	8
070 - Fatty acid biosynthesis	2.3.1	2.3.1	1.1.1	4.2.1	1.3.1	2.3.1	1.1.1	4.2.1	1.3.1
046 - Pentose and glucuronate interconversions	5.3.1	-.-.-	1.1.1	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	-.-.-
196 - Valine, leucine and isoleucine biosynthesis	4.2.1	-.-.-	-.-.-	4.2.1	1.1.1	2.2.1	1.1.1	-.-.-	-.-.-
213 - Histidine metabolism	-.-.-	-.-.-	-.-.-	4.2.1	1.1.1	-.-.-	1.1.1	-.-.-	6.1.1
073 - Fatty acid biosynthesis	-.-.-	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.3.1	-.-.-	-.-.-
044 - Pentose and glucuronate interconversions	-.-.-	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	-.-.-
054 - Fructose and mannose metabolism	2.7.7	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	-.-.-
212 - Histidine metabolism	-.-.-	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	6.1.1
021 - Citrate cycle (TCA cycle)	1.1.1	2.3.3	-.-.-	4.2.1	-.-.-	4.2.1	1.1.1	-.-.-	1.1.1

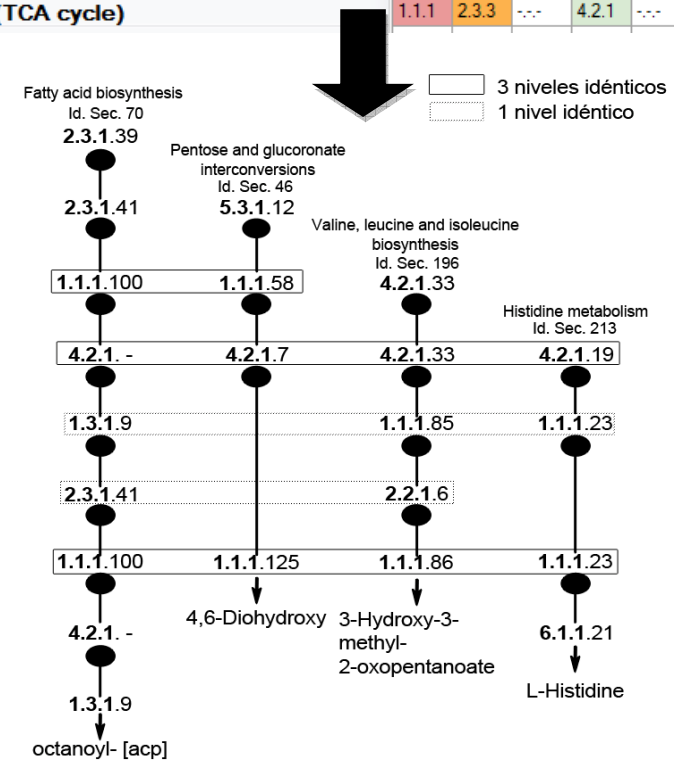


Figura 5.9. Mapeo de las similitudes encontradas en el alineamiento del *cluster* 33 en las vías metabólicas.

En la Figura 5.9 se muestra un ejemplo de este mapeo. En la parte superior tenemos el alineamiento del *cluster* 33 obtenido con el AG progresivo y la función de entropía. Las columnas 3 y 6 son idénticas en las primeras cuatro secuencias del *cluster*. En la parte inferior se tienen las 4 secuencias (reacciones enzimáticas) de la vía desde la enzima de inicio hasta el producto final de cada secuencia, representando las enzimas con los 4 niveles de los números enzimáticos. Las primeras dos vías pertenecen al metabolismo de los lípidos y al de los carbohidratos respectivamente, y las últimas dos al de los aminoácidos. En un principio no se esperaría alguna relación entre estos tres diferentes metabolismos, sin embargo encontramos pasos intermedios en la primera vía similares a las otras tres. En el caso de la síntesis de la valina, leucina e isoleucina, se encuentran los cuatro pasos, el primero y el último, idénticos en los primeros tres niveles, y los intermedios similares en el primer nivel. Esto indica que las enzimas presentes en estos segmentos de las vías tienen funciones catalíticas similares, lo que sugiere que existen reacciones similares en la síntesis de estos aminoácidos y en la parte intermedia de la síntesis del octanoilo. Lo mismo podríamos decir de la síntesis de la histidina, pero se puede observar que en este caso hay un paso que no está presente, esto podría indicar que el mecanismo para esta última vía no requiere de ese tipo de reacción presente en la primera. Resultados similares se pueden encontrar analizando los alineamientos de los demás *clusters*.

Capítulo 6

CONCLUSIONES

La transformación de las reacciones enzimáticas asociadas a una ruta metabólica en secuencias de reacciones nos permitió implementar el enfoque del alineamiento de secuencias de aminoácidos en el algoritmo propuesto para el alineamiento múltiple de vías metabólicas (reacciones). Esta representación permite visualizar de forma fácil y eficiente las similitudes y diferencias entre regiones completas de vías metabólicas, lo cual es novedoso en el análisis comparativo y evolutivo del metabolismo.

El algoritmo genético (AG) propuesto resultó ser altamente eficaz al combinarse con la técnica de alineamientos progresivos. En este contexto, el uso de los alineamientos en pares, como parte del algoritmo, permitió dirigir de mejor manera la búsqueda. Al incluir información relacionada al problema dentro de los operadores, los algoritmos genéticos demuestran un mejor desempeño.

La función objetivo tiene un gran peso dentro del algoritmo, ya que es la principal guía que usa para explorar el espacio de búsqueda. Las dos funciones propuestas obtuvieron resultados similares, logrando minimizar el número de eventos (*gaps*) y maximizando la homogeneidad dentro de las columnas alineadas.

El algoritmo de agrupamiento o *clustering*, propuesto como parte del análisis evolutivo de las secuencias de reacciones, fue de gran utilidad para identificar las posibles similitudes dentro de vías pertenecientes a metabolismos diferentes, como se muestra en la Figura 5.9 del capítulo anterior, donde se encontraron similitudes en tres diferentes metabolismos, o en los *clusters* 2 y 30 (ver apéndice E) donde tenemos 8 diferentes metabolismos que comparten al menos dos pasos catalíticos. Los grupos que resultaron de este algoritmo permitieron realizar alineamientos múltiples de mayor calidad y que evidencian las regiones conservadas entre las vías metabólicas.

La diferencia principal con los métodos existentes está en la comparación de vías metabólicas a un nivel funcional, considerando la clasificación de las enzimas que

intervienen en ellas, sin tomar en cuenta las secuencias de aminoácidos que representan a estas enzimas. En este contexto, el enfoque propuesto en este trabajo resulta ser altamente innovador.

Los resultados obtenidos, encontrando módulos de enzimas repetidas al interior de una misma vía metabólica, así como en vías de metabolismos diferentes como son el de los carbohidratos y el de los aminoácidos, permiten apoyar resultados presentados en trabajos previos acerca de las teorías evolutivas de las vías metabólicas. La idea de que la duplicación de genes es importante en la generación de variantes e innovaciones dentro de las vías metabólicas se comprueba al encontrar estos segmentos. Adicionalmente, se identificó que estos pasos metabólicos repetidos tienen restricciones en el orden en el que aparecen. Es decir, ciertos tipos de enzimas (pasos catalíticos) usualmente aparecen seguidas de otro tipo (se observa comparando los números enzimáticos del primer y segundo nivel), así como en el tamaño (número de enzimas) [18].

El enfoque del alineamiento de secuencias empleado en el análisis evolutivo de proteínas resultó también ser efectivo en el análisis de las vías metabólicas desde una perspectiva independiente de las secuencias de aminoácidos.

Capítulo 7

PERSPECTIVAS

El algoritmo genético implementado en este trabajo es una primera aproximación del cómputo evolutivo para resolver el problema del alineamiento de vías metabólicas. Es posible realizar mejoras gracias a la facilidad de adaptación que tienen este tipo de algoritmos. Crear nuevos operadores que permitan explorar el espacio de búsqueda de una mejor manera, y usar una representación diferente integrando otros elementos de las vías metabólicas son posibles cambios que podrían mejorar el desempeño del algoritmo tanto en tiempo como en la calidad de los resultados. Una representación que no requiera correcciones para asegurar el tamaño adecuado de la matriz y a su vez elimine este problema en los operadores de cruza y mutación es una de las áreas en las que se podría trabajar más. Otro punto importante a revisar es la función objetivo que debe integrar la mayor información posible sobre el problema para obtener una búsqueda mucho más eficiente. Las dos funciones propuestas obtuvieron buenos resultados, pero aún es posible realizar un mejor análisis para ajustar los parámetros que estas funciones requieren.

Probar otros algoritmos de *clustering* (posiblemente mapas auto-organizados) podría proporcionar una manera de validar los resultados que se obtuvieron con el algoritmo *K-Means* ya que, como se mencionó anteriormente, el elegir el número de grupos existentes es un problema importante para este tipo de algoritmos.

En este trabajo se realizó el análisis de las vías metabólicas de un sólo organismo, pero contando con esta herramienta es posible ahora llevar a cabo la comparación de todas las vías reportadas en la base de datos KEGG, lo cual permitirá responder preguntas sobre el origen evolutivo del metabolismo de diferentes organismos, así como dar indicios de rutas alternas dentro de las vías desde una perspectiva más integral.

Apéndice A

Relación de las 47 vías metabólicas de *E. coli* y el número de secuencias que se encontró en cada una.

# mapa	ID	Nombre	Tipo Metabolismo	Num. Secs.	Long. Promedio
1	eco00010	Glycolysis / Gluconeogenesis	Carbohydrate Metabolism	20	5.25
2	eco00020	Citrate cycle (TCA cycle)	Carbohydrate Metabolism	4	3.75
3	eco00030	Pentose phosphate pathway	Carbohydrate Metabolism	12	3.42
4	eco00040	Pentose and glucuronate interconversions	Carbohydrate Metabolism	12	2.83
5	eco00051	Fructose and mannose metabolism	Carbohydrate Metabolism	9	3.44
6	eco00052	Galactose metabolism	Carbohydrate Metabolism	8	3.00
7	eco00053	Ascorbate and aldarate metabolism	Carbohydrate Metabolism	3	3.67
8	eco00061	Fatty acid biosynthesis	Lipid Metabolism	5	4.00
9	eco00071	Fatty acid metabolism	Lipid Metabolism	1	2.00
10	eco00100	Biosynthesis of steroids	Lipid Metabolism	2	8.00
11	eco00130	Ubiquinone and menaquinone biosynthesis	Metabolism of Cofactors and Vitamins	3	5.67
12	eco00220	Urea cycle and metabolism of amino groups	Amino Acid Metabolism	6	4.33
13	eco00240	Pyrimidine metabolism	Nucleotide Metabolism	34	8.03
14	eco00251	Glutamate metabolism	Amino Acid Metabolism	39	5.00
15	eco00252	Alanine and aspartate metabolism	Amino Acid Metabolism	1	2.00
16	eco00260	Glycine, serine and threonine metabolism	Amino Acid Metabolism	16	3.38
17	eco00271	Methionine metabolism	Amino Acid Metabolism	13	7.15
18	eco00272	Cysteine metabolism	Amino Acid Metabolism	4	2.75
19	eco00280	Valine, leucine and isoleucine degradation	Amino Acid Metabolism	1	4.00
20	eco00290	Valine, leucine and isoleucine biosynthesis	Amino Acid Metabolism	6	3.33
21	eco00300	Lysine biosynthesis	Amino Acid Metabolism	4	3.25
22	eco00310	Lysine degradation	Amino Acid Metabolism	2	2.50
23	eco00330	Arginine and proline metabolism	Amino Acid Metabolism	5	3.80
24	eco00340	Histidine metabolism	Amino Acid Metabolism	5	3.40
25	eco00360	Phenylalanine metabolism	Amino Acid Metabolism	6	4.00
26	eco00362	Benzoate degradation via hydroxylation	Xenobiotics Biodegradation and Metabolism	1	2.00
27	eco00380	Tryptophan metabolism	Amino Acid Metabolism	1	3.00
28	eco00400	Phenylalanine, tyrosine and tryptophan biosynthesis	Amino Acid Metabolism	8	4.38
29	eco00410	Beta-Alanine metabolism	Metabolism of Other Amino	4	2.00

			Acids		
30	eco00430	Taurine and hypotaurine metabolism	Metabolism of Other Amino Acids	2	2.00
31	eco00450	Selenoamino acid metabolism	Metabolism of Other Amino Acids	3	2.33
32	eco00471	D-Glutamine and D-glutamate metabolism	Metabolism of Other Amino Acids	2	2.50
33	eco00473	D-Alanine metabolism	Metabolism of Other Amino Acids	1	2.00
34	eco00480	Glutathione metabolism	Metabolism of Other Amino Acids	5	4.60
35	eco00500	Starch and sucrose metabolism	Carbohydrate Metabolism	16	6.88
36	eco00520	Nucleotide sugars metabolism	Carbohydrate Metabolism	3	4.00
37	eco00521	Streptomycin biosynthesis	Biosynthesis of Other Secondary Metabolites	1	3.00
38	eco00530	Aminosugars metabolism	Carbohydrate Metabolism	20	7.70
39	eco00540	Lipopolysaccharide biosynthesis	Glycan Biosynthesis and Metabolism	3	6.00
40	eco00550	Peptidoglycan biosynthesis	Glycan Biosynthesis and Metabolism	3	4.67
41	eco00561	Glycerolipid metabolism	Lipid Metabolism	2	3.50
42	eco00564	Glycerophospholipid metabolism	Lipid Metabolism	7	4.00
43	eco00603	Glycosphingolipid biosynthesis - globo series	Glycan Biosynthesis and Metabolism	1	2.00
44	eco00604	Glycosphingolipid biosynthesis - ganglio series	Glycan Biosynthesis and Metabolism	1	2.00
45	eco00620	Pyruvate metabolism	Carbohydrate Metabolism	145	8.96
46	eco00621	Biphenyl degradation	Xenobiotics Biodegradation and Metabolism	1	2.00
47	eco00629	Carbazole degradation	Xenobiotics Biodegradation and Metabolism	1	2.00

Apéndice B

Relación de las 452 secuencias que se encontraron, el identificador, el *cluster* y la vía a la que pertenecen.

Cluster	ID Sec.	Vía	Secuencia
1	133	eco00251	1.5.99:1.4.1:1.4.1:4.1.1:2.6.1:1.2.1
1	209	eco00330	1.5.99:1.5.99
1	150	eco00251	1.5.99:6.3.1:1.4.1:1.4.1:1.4.1:4.1.1:2.6.1:1.2.1
1	147	eco00251	1.5.99:6.3.1:1.4.1:1.4.1:2.6.1:4.1.1:2.6.1:1.2.1
1	125	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:1.4.1:4.1.1:2.6.1:1.2.1
1	144	eco00251	1.5.99:4.1.1:2.6.1:1.2.1
1	80	eco00220	2.6.1:1.2.1
1	83	eco00220	3.5.1:4.1.1:2.6.1:1.2.1
1	122	eco00251	1.5.99:1.4.1:2.6.1:4.1.1:2.6.1:1.2.1
1	153	eco00251	1.5.99:6.3.1:1.4.1:4.1.1:2.6.1:1.2.1
1	136	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:1.4.1:4.1.1:2.6.1:1.2.1
2	38	eco00040	2.7.1:5.1.3
2	39	eco00040	2.7.1:5.-.-:5.1.3
2	289	eco00540	5.-.-:2.7.7:3.1.3:2.7.7:5.1.3
2	265	eco00520	2.7.7:5.1.3
2	268	eco00521	2.7.7:4.2.1:5.1.3
2	29	eco00030	2.7.1:5.3.1:5.1.3
2	42	eco00040	2.7.1:4.1.1:5.-.-:5.1.3
2	64	eco00052	2.7.7:5.1.3
2	66	eco00053	2.7.1:3.1.1:4.1.1:5.1.3:5.1.3
2	269	eco00530	3.2.1:2.7.1:5.1.3
2	266	eco00520	2.7.7:4.2.1:5.1.3:1.1.1
2	229	eco00400	2.7.1:2.5.1:4.2.3:5.4.99:5.4.99
2	278	eco00530	2.7.1:4.2.-:5.1.3
2	67	eco00053	2.7.1:4.1.1:5.1.3:5.1.3
2	32	eco00030	2.7.1:1.1.1:5.1.3
3	432	eco00620	1.1.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
3	361	eco00620	3.1.2:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
3	324	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
3	397	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
3	22	eco00020	1.1.1:2.3.3:4.1.3:4.1.1
3	343	eco00620	4.4.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
3	433	eco00620	1.1.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
3	326	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
3	396	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1

3	434	eco00620	1.1.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
3	378	eco00620	2.3.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
3	306	eco00620	4.2.3:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
3	194	eco00290	4.2.1:2.3.3:4.2.1
3	398	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
3	360	eco00620	3.1.2:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
3	342	eco00620	4.4.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
3	308	eco00620	4.2.3:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
3	362	eco00620	3.1.2:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
3	325	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
3	379	eco00620	2.3.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
3	380	eco00620	2.3.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
3	236	eco00430	4.1.1:1.14.11
3	307	eco00620	4.2.3:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
3	344	eco00620	4.4.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
4	240	eco00450	2.7.7:2.7.1
4	76	eco00100	2.2.1:1.1.1:2.7.7:2.7.1:4.6.1:1.17.7:1.17.1:2.5.1
4	111	eco00240	2.7.4:2.7.4:2.7.7
4	28	eco00030	2.7.1:2.7.6
4	63	eco00052	3.2.1:2.7.1:2.7.7
4	295	eco00561	3.2.1:2.7.1:2.3.1:2.3.1
4	75	eco00100	2.2.1:1.1.1:2.7.7:2.7.1:4.6.1:1.17.7:1.17.1:5.3.3
4	299	eco00564	3.1.4:2.3.1:2.3.1:2.7.7:2.7.8:3.1.1
4	291	eco00540	2.3.1:3.5.1:2.3.1:3.6.1:2.4.1:2.7.1:2.-.-:2.-.-:2.3.1:2.3.1
4	297	eco00564	2.7.1:2.7.7:2.7.8:3.1.1
4	59	eco00052	3.2.1:3.2.1:2.7.1:2.7.7
4	300	eco00564	3.1.4:2.3.1:2.3.1:2.7.7:2.7.8:3.1.3:2.7.8
5	280	eco00530	2.7.1:4.2.-:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:5.1.3:1.1.1
5	283	eco00530	2.7.1:4.2.-:3.5.1:5.4.2:2.7.7:2.7.7:2.5.1:1.1.1
5	285	eco00530	2.7.1:4.2.-:3.5.1:5.4.2:2.7.7:2.7.7:5.1.3:3.1.3
5	3	eco00010	5.4.2:2.7.2
5	275	eco00530	3.2.1:2.7.1:3.5.1:5.4.2:2.7.7:2.7.7:5.1.3:1.1.1
5	281	eco00530	2.7.1:4.2.-:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:5.1.3:3.1.3
5	272	eco00530	3.2.1:2.7.1:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:5.1.3:3.1.3
5	286	eco00530	2.7.1:4.2.-:3.5.1:5.4.2:2.7.7:2.7.7:5.1.3:2.7.1
5	282	eco00530	2.7.1:4.2.-:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:5.1.3:2.7.1
5	204	eco00310	4.2.1:5.1.2:2.3.1
5	26	eco00030	5.4.2:2.7.6
5	270	eco00530	3.2.1:2.7.1:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:2.5.1:1.1.1
5	273	eco00530	3.2.1:2.7.1:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:5.1.3:2.7.1
5	223	eco00380	4.2.1:5.1.2:2.3.1
5	264	eco00500	2.4.1:5.4.2:5.3.1

5	271	eco00530	3.2.1:2.7.1:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:5.1.3:1.1.1
5	277	eco00530	3.2.1:2.7.1:3.5.1:5.4.2:2.7.7:2.7.7:5.1.3:2.7.1
5	193	eco00280	4.2.1:5.1.2:2.3.1:2.3.1
5	279	eco00530	2.7.1:4.2.-:3.5.1:3.5.99:2.6.1:5.4.2:2.7.7:2.7.7:2.5.1:1.1.1
5	276	eco00530	3.2.1:2.7.1:3.5.1:5.4.2:2.7.7:2.7.7:5.1.3:3.1.3
5	274	eco00530	3.2.1:2.7.1:3.5.1:5.4.2:2.7.7:2.7.7:2.5.1:1.1.1
5	284	eco00530	2.7.1:4.2.-:3.5.1:5.4.2:2.7.7:2.7.7:5.1.3:1.1.1
6	152	eco00251	1.5.99:6.3.1:1.4.1:6.1.1
6	155	eco00251	1.5.99:6.3.1:2.4.2
6	151	eco00251	1.5.99:6.3.1:1.4.1:5.1.1
6	157	eco00251	1.5.99:6.3.1:3.5.1
6	148	eco00251	1.5.99:6.3.1:1.4.1:1.4.1:1.4.1:5.1.1
6	145	eco00251	1.5.99:6.3.1:1.4.1:1.4.1:2.6.1:5.1.1
6	143	eco00251	1.5.99:6.1.1
6	158	eco00251	1.5.99:6.3.1:6.1.1
6	146	eco00251	1.5.99:6.3.1:1.4.1:1.4.1:2.6.1:6.1.1
6	156	eco00251	1.5.99:6.3.1:6.3.5
6	142	eco00251	1.5.99:5.1.1
6	149	eco00251	1.5.99:6.3.1:1.4.1:1.4.1:1.4.1:6.1.1
7	441	eco00620	1.1.1:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	389	eco00620	2.3.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
7	387	eco00620	2.3.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	351	eco00620	4.4.1:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	393	eco00620	2.3.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	370	eco00620	3.1.2:1.1.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
7	371	eco00620	3.1.2:1.1.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
7	395	eco00620	2.3.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
7	333	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	369	eco00620	3.1.2:1.1.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	335	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
7	363	eco00620	3.1.2:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	352	eco00620	4.4.1:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
7	316	eco00620	4.2.3:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
7	61	eco00052	9.9.9:1.1.1:2.7.1
7	353	eco00620	4.4.1:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
7	317	eco00620	4.2.3:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
7	315	eco00620	4.2.3:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
7	334	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
7	394	eco00620	2.3.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
7	442	eco00620	1.1.1:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
7	388	eco00620	2.3.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
7	443	eco00620	1.1.1:1.2.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1

8	405	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
8	403	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
8	385	eco00620	2.3.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
8	406	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
8	411	eco00620	1.2.1:2.3.3:1.1.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
8	400	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
8	404	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
8	401	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
8	412	eco00620	1.2.1:2.3.3:1.1.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
8	399	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
8	402	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
8	407	eco00620	1.2.1:2.3.3:1.1.1:2.7.9:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
8	413	eco00620	1.2.1:2.3.3:1.1.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
8	74	eco00071	6.2.1:1.3.99
9	219	eco00360	1.14.12:1.3.1:1.13.11:3.7.1:4.2.1:4.1.3
9	165	eco00260	3.1.3:4.3.1
9	216	eco00360	1.14.13:1.13.11:3.7.1:4.2.1:4.1.3
9	35	eco00030	5.3.1:1.1.1:3.1.1:4.2.1:4.1.3:2.2.1
9	231	eco00400	1.1.1:4.2.1
9	23	eco00020	1.1.1:4.1.1
9	221	eco00360	1.14.12:1.3.1:1.13.11:3.7.1
9	220	eco00360	1.14.13:1.13.11:3.7.1
10	161	eco00260	2.7.1:1.1.1
10	252	eco00500	2.4.1:2.7.7:1.1.1
10	41	eco00040	2.7.7:1.1.1
10	259	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:2.4.1:2.7.7:1.1.1
10	71	eco00061	2.3.1:2.3.1:1.1.1:4.2.1:1.3.1
10	199	eco00290	4.3.1:2.2.1:1.1.1
10	198	eco00290	2.2.1:2.2.1:1.1.1
10	267	eco00520	2.7.7:1.1.1:2.1.2:2.6.1:2.1.2:2.7.8
11	113	eco00240	3.1.3:3.5.4:2.7.1:2.7.4:2.7.4:3.6.1:2.1.1:3.1.3:2.7.1:2.7.4:2.7.4:2.7.7
11	109	eco00240	2.7.4:2.7.4:3.5.4:3.6.1:2.1.1:3.1.3:2.7.1:2.7.4:2.7.4:2.7.7
11	110	eco00240	2.7.4:2.7.4:3.5.4:3.6.1:2.1.1:3.1.3:2.4.2
11	112	eco00240	3.1.3:3.5.4:2.7.1:2.7.4:2.7.4:3.6.1:2.1.1:2.7.4:2.7.4:2.7.7
11	116	eco00240	3.1.3:3.5.4:2.7.1:2.1.1:3.1.3:2.7.1:2.7.4:2.7.4:2.7.7
11	298	eco00564	2.7.1:2.7.7:2.7.8:3.1.3:2.7.8
11	107	eco00240	2.7.4:2.7.4:3.5.4:3.6.1:2.7.4:2.7.4:3.6.1:2.1.1:3.1.3:2.4.2
11	105	eco00240	2.7.4:2.7.4:3.5.4:3.6.1:2.7.4:2.7.4:3.6.1:2.1.1:2.7.4:2.7.4:2.7.7
11	106	eco00240	2.7.4:2.7.4:3.5.4:3.6.1:2.7.4:2.7.4:3.6.1:2.1.1:3.1.3:2.7.1:2.7.4:2.7.4:2.7.7
11	108	eco00240	2.7.4:2.7.4:3.5.4:3.6.1:2.1.1:2.7.4:2.7.4:2.7.7
11	117	eco00240	3.1.3:3.5.4:2.7.1:2.1.1:3.1.3:2.4.2
11	114	eco00240	3.1.3:3.5.4:2.7.1:2.7.4:2.7.4:3.6.1:2.1.1:3.1.3:2.4.2

11	290	eco00540	2.5.1:3.1.3:2.7.7
11	115	eco00240	3.1.3:3.5.4:2.7.1:2.1.1:2.7.4:2.7.4:2.7.7
12	121	eco00251	1.5.99:1.4.1:2.6.1:6.1.1
12	127	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:2.4.2
12	154	eco00251	1.5.99:6.3.1:2.6.1
12	208	eco00330	1.5.99:2.6.1:4.1.3
12	218	eco00360	1.4.99:2.6.1
12	130	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:6.1.1
12	120	eco00251	1.5.99:1.4.1:2.6.1:5.1.1
12	129	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:3.5.1
12	123	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:1.4.1:5.1.1
12	128	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:6.3.5
12	126	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:2.6.1
13	435	eco00620	1.1.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
13	347	eco00620	4.4.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
13	436	eco00620	1.1.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
13	437	eco00620	1.1.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
13	53	eco00051	2.7.7:9.9.9
13	327	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
13	329	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
13	346	eco00620	4.4.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
13	311	eco00620	4.2.3:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
13	383	eco00620	2.3.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
13	119	eco00240	6.3.5:9.9.9
13	345	eco00620	4.4.1:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
13	365	eco00620	3.1.2:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
13	309	eco00620	4.2.3:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
13	328	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
13	381	eco00620	2.3.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
13	310	eco00620	4.2.3:1.2.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
13	382	eco00620	2.3.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
13	364	eco00620	3.1.2:1.1.1:2.7.9:4.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
14	6	eco00010	2.7.1:5.3.1:5.3.1:2.7.1:4.1.2:1.2.1
14	47	eco00040	3.2.1:5.3.1
14	10	eco00010	5.4.2:5.3.1:5.3.1:2.7.1:4.1.2:1.2.1
14	224	eco00400	5.3.1:5.3.1:4.2.1
14	19	eco00010	2.7.1:3.2.1:5.3.1:2.7.1:4.1.2:1.2.1
14	4	eco00010	2.7.1:5.3.1:2.7.1:4.1.2:1.2.1
14	8	eco00010	5.4.2:5.3.1:2.7.1:4.1.2:1.2.1
14	225	eco00400	5.3.1:5.3.1:4.2.1:4.2.1
14	214	eco00340	3.6.1:3.6.1:5.3.1:4.1.3
14	16	eco00010	3.1.3:2.7.1:5.3.1:5.3.1:2.7.1:4.1.2:1.2.1

14	27	eco00030	5.4.2:5.3.1:5.1.3
15	446	eco00620	1.1.1:1.2.1:1.2.2:6.2.1:6.2.1:6.4.1
15	356	eco00620	4.4.1:1.2.1:1.2.2:6.2.1:6.2.1:6.4.1
15	392	eco00620	2.3.1:1.2.2:6.2.1:6.2.1:6.4.1
15	72	eco00061	6.4.1:6.4.1
15	338	eco00620	1.1.1:1.2.1:1.1.2:1.2.2:6.2.1:6.2.1:6.4.1
15	332	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
15	320	eco00620	4.2.3:1.2.1:1.2.2:6.2.1:6.2.1:6.4.1
15	449	eco00620	1.1.1:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
15	341	eco00620	1.1.1:1.2.1:1.1.2:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
15	374	eco00620	3.1.2:1.1.1:1.2.2:6.2.1:6.2.1:6.4.1
15	431	eco00620	1.2.1:6.4.1
16	211	eco00340	3.4.13:6.1.1
16	235	eco00410	3.4.13:6.3.2
16	163	eco00260	3.1.3:6.1.1
17	251	eco00500	2.4.1:2.7.7:2.4.1:3.2.1:3.2.1
17	250	eco00500	2.4.1:2.7.7:2.4.1:3.2.1
17	262	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:2.4.1:2.7.7:2.4.1:3.1.3:2.7.1:3.2.1
17	62	eco00052	3.2.1:3.2.1
17	257	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:2.4.1:2.7.7:2.4.1:3.2.1
17	253	eco00500	2.4.1:2.7.7:2.4.1:3.1.3:3.2.1:2.4.1:3.2.1:2.7.1:5.3.1
17	256	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:3.2.1
17	258	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:2.4.1:2.7.7:2.4.1:3.2.1:3.2.1
17	58	eco00052	3.2.1:3.2.1:3.2.1
17	255	eco00500	2.4.1:2.7.7:2.4.1:3.1.3:2.7.1:3.2.1
17	261	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:2.4.1:2.7.7:2.4.1:3.1.3:3.2.1:2.4.1:2.7.1
17	304	eco00603	3.2.1:3.2.1
17	305	eco00604	3.2.1:3.2.1
17	260	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:2.4.1:2.7.7:2.4.1:3.1.3:3.2.1:2.4.1:3.2.1:2.7.1:5.3.1
17	263	eco00500	2.4.1:2.7.7:2.4.1:2.4.1:2.4.1:5.4.2:5.3.1
17	254	eco00500	2.4.1:2.7.7:2.4.1:3.1.3:3.2.1:2.4.1:2.7.1
18	81	eco00220	3.5.1:2.1.3:6.3.4:4.3.2:4.1.1:3.5.3:2.6.1:1.2.1
18	82	eco00220	3.5.1:2.1.3:6.3.4:4.3.2:4.1.1:3.5.3:2.3.1
18	234	eco00410	3.4.13:2.6.1
18	207	eco00330	2.1.3:6.3.4:4.3.2:6.1.1
18	206	eco00330	2.1.3:6.3.4:4.3.2:2.3.1:3.5.3:2.6.1:1.2.1:3.5.1
18	159	eco00252	3.4.13:2.6.1
19	132	eco00251	1.5.99:1.4.1:1.4.1:6.1.1
19	137	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:2.6.1
19	141	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:6.1.1
19	134	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:1.4.1:5.1.1
19	124	eco00251	1.5.99:1.4.1:2.6.1:6.3.1:1.4.1:6.1.1

19	138	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:2.4.2
19	140	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:3.5.1
19	131	eco00251	1.5.99:1.4.1:1.4.1:5.1.1
19	135	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:1.4.1:6.1.1
19	139	eco00251	1.5.99:1.4.1:1.4.1:6.3.1:6.3.5
19	210	eco00330	1.5.1:1.5.99
20	438	eco00620	1.1.1:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
20	312	eco00620	4.2.3:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
20	391	eco00620	2.3.1:1.2.2:6.2.1:6.2.1:2.3.3
20	313	eco00620	4.2.3:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
20	439	eco00620	1.1.1:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
20	367	eco00620	3.1.2:1.1.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
20	368	eco00620	3.1.2:1.1.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
20	366	eco00620	3.1.2:1.1.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
20	386	eco00620	2.3.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
20	314	eco00620	4.2.3:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
20	440	eco00620	1.1.1:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
20	200	eco00300	2.7.2:1.2.1:4.2.1
20	330	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
20	162	eco00260	2.7.2:1.2.1
20	384	eco00620	2.3.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
20	331	eco00620	1.1.1:1.2.1:1.1.2:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
20	390	eco00620	2.3.1:1.2.2:6.2.1:6.2.1:2.3.1
20	349	eco00620	4.4.1:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
20	348	eco00620	4.4.1:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
20	350	eco00620	4.4.1:1.2.1:2.7.9:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
21	15	eco00010	3.1.3:2.7.1:5.3.1:2.7.1:4.1.2:5.3.1
21	9	eco00010	5.4.2:5.3.1:2.7.1:4.1.2:5.3.1
21	40	eco00040	5.3.1:2.7.1
21	52	eco00051	5.4.2:5.3.1:2.7.1:4.1.2
21	13	eco00010	3.1.3:5.1.3:2.7.1:5.3.1:2.7.1:4.1.2:5.3.1
21	20	eco00010	2.7.1:3.2.1:5.3.1:2.7.1:4.1.2:5.3.1
21	30	eco00030	5.3.1:2.7.1:4.1.2:2.2.1
21	17	eco00010	3.1.3:2.7.1:5.3.1:5.3.1:2.7.1:4.1.2:5.3.1
21	57	eco00051	5.3.1:2.7.1:2.7.1:4.1.2
21	11	eco00010	5.4.2:5.3.1:5.3.1:2.7.1:4.1.2:5.3.1
21	60	eco00052	2.7.1:5.3.1:2.7.1
21	12	eco00010	3.1.3:5.1.3:2.7.1:5.3.1:2.7.1:4.1.2:1.2.1
21	14	eco00010	3.1.3:2.7.1:5.3.1:2.7.1:4.1.2:1.2.1
21	7	eco00010	2.7.1:5.3.1:5.3.1:2.7.1:4.1.2:5.3.1
21	37	eco00040	5.3.1:2.7.1:5.1.3
21	5	eco00010	2.7.1:5.3.1:2.7.1:4.1.2:5.3.1

21	56	eco00051	5.3.1:2.7.1:4.1.2
21	249	eco00500	3.2.1:2.7.1:5.3.1
22	1	eco00010	4.1.1:2.7.1
22	232	eco00410	4.1.1:2.6.1
22	77	eco00130	4.1.3:2.5.1:4.1.1:9.9.9:2.1.1:1.14.13:2.1.1:1.14.13:2.1.1
22	2	eco00010	4.2.1:2.7.1
22	98	eco00240	4.1.1:2.4.2
22	288	eco00530	4.1.3:2.7.1
22	25	eco00030	5.4.2:4.1.2:2.2.1
22	85	eco00220	4.1.1:2.5.1
22	237	eco00430	4.1.1:2.3.2
22	84	eco00220	3.5.1:4.1.1:2.3.1
23	301	eco00564	3.1.4:1.1.1
23	296	eco00561	3.2.1:1.1.1:2.7.1
23	34	eco00030	5.3.1:1.1.1:3.1.1:1.1.1:5.1.3
23	302	eco00564	3.1.4:1.1.5
24	242	eco00471	3.5.1:5.1.1:6.3.2
24	203	eco00300	3.5.1:5.1.1:6.3.2:6.3.2
24	233	eco00410	4.1.1:6.3.2
24	243	eco00473	5.1.1:6.3.2
24	241	eco00471	3.5.1:6.3.2
24	202	eco00300	3.5.1:5.1.1:4.1.1:6.1.1
25	188	eco00271	2.3.1:2.5.1:2.1.1:2.5.1:2.1.1:3.2.2:4.4.1:2.1.1:6.1.1:2.1.2
25	180	eco00271	2.3.1:2.5.1:4.4.1:2.1.1:2.5.1:4.1.1:2.5.1:3.2.2
25	181	eco00271	2.3.1:2.5.1:4.4.1:2.1.1:2.5.1:2.1.1:3.2.2:4.4.1:2.1.1:6.1.1:2.1.2
25	179	eco00271	2.3.1:2.5.1:4.4.1:2.1.1:6.1.1:2.1.2
25	178	eco00271	2.6.1:2.5.1:2.1.1:3.2.2:4.4.1:2.1.1:6.1.1:2.1.2
25	177	eco00271	2.6.1:2.5.1:4.1.1:2.5.1:3.2.2
25	184	eco00271	2.3.1:2.5.1:2.5.1:2.5.1:2.1.1:2.5.1:2.1.1:3.2.2:4.4.1:2.1.1:6.1.1:2.1.2
25	239	eco00450	2.5.1:2.5.1:4.4.1
25	185	eco00271	2.3.1:2.5.1
25	191	eco00272	2.3.1:2.5.1:6.1.1
25	183	eco00271	2.3.1:2.5.1:2.5.1:2.5.1:2.1.1:2.5.1:4.1.1:2.5.1:3.2.2
25	176	eco00271	2.6.1:6.1.1:2.1.2
25	79	eco00130	2.5.1:2.1.1
25	69	eco00061	2.3.1:2.3.1
25	186	eco00271	2.3.1:2.5.1:2.1.1:6.1.1:2.1.2
25	182	eco00271	2.3.1:2.5.1:2.5.1:2.5.1:2.1.1:6.1.1:2.1.2
25	238	eco00450	2.5.1:4.4.1
25	187	eco00271	2.3.1:2.5.1:2.1.1:2.5.1:4.1.1:2.5.1:3.2.2
26	247	eco00480	2.5.1:2.3.2:6.3.2:6.3.2:2.3.2:3.4.11
26	244	eco00480	2.3.2:3.4.11

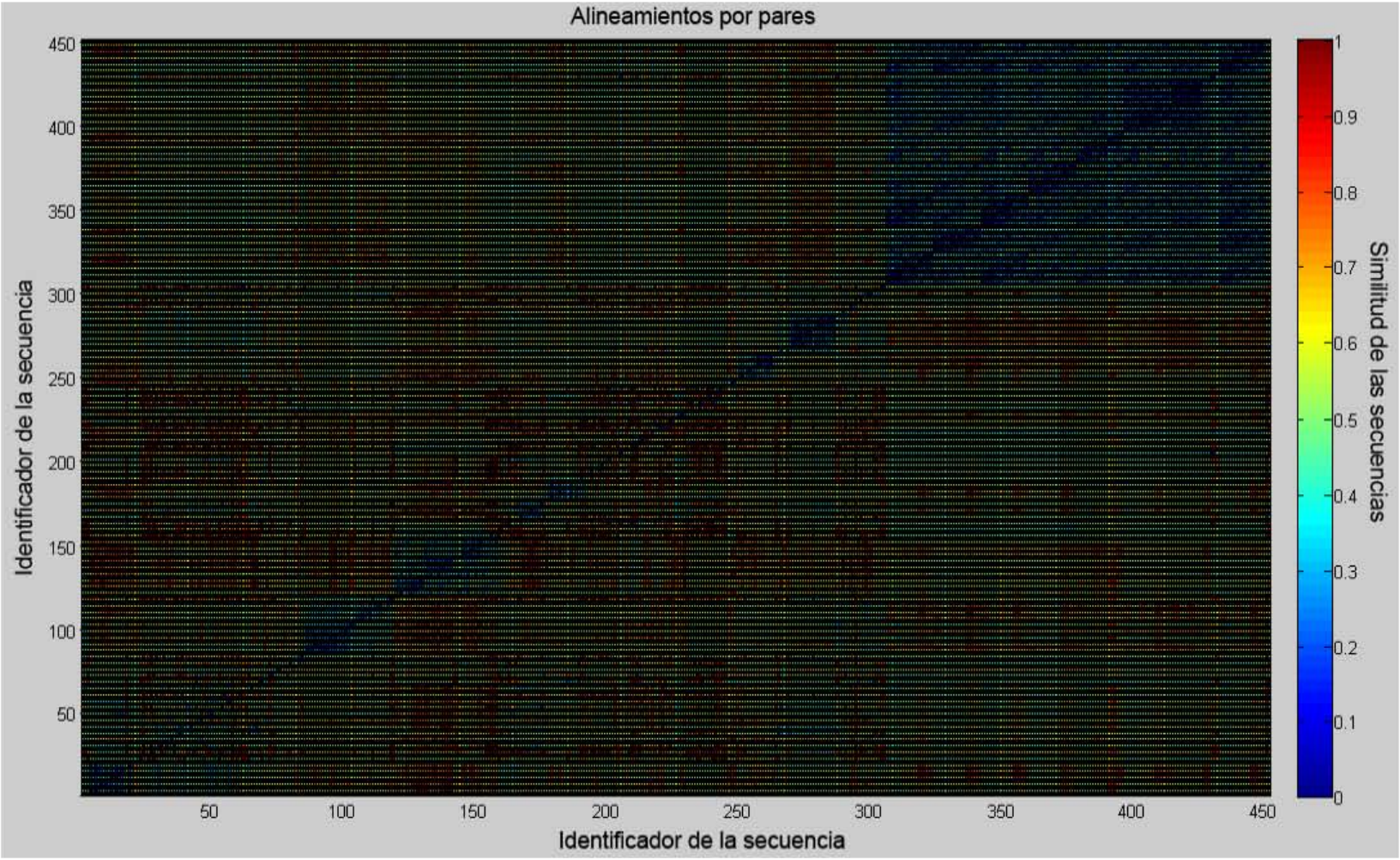
26	245	eco00480	2.3.2:3.4.13
26	248	eco00480	2.5.1:2.3.2:6.3.2:6.3.2:2.3.2:3.4.13
26	292	eco00550	6.3.2:6.3.2:2.7.8:2.4.1:6.3.1
26	293	eco00550	6.3.2:6.3.2:2.7.8:2.4.1:2.4.1:3.6.1
26	294	eco00550	6.3.2:6.3.2:6.3.2
26	246	eco00480	2.5.1:2.3.2:6.3.2:6.3.2:1.11.1:1.8.1:1.1.1
27	447	eco00620	1.1.1:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
27	321	eco00620	4.2.3:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
27	322	eco00620	4.2.3:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
27	375	eco00620	3.1.2:1.1.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
27	303	eco00564	3.1.4:2.3.1
27	358	eco00620	4.4.1:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
27	376	eco00620	3.1.2:1.1.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
27	448	eco00620	1.1.1:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
27	323	eco00620	4.2.3:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
27	339	eco00620	1.1.1:1.2.1:1.1.2:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
27	426	eco00620	1.2.1:2.3.1:3.6.1:6.2.1:6.2.1:2.3.1
27	340	eco00620	1.1.1:1.2.1:1.1.2:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
27	359	eco00620	4.4.1:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
27	357	eco00620	4.4.1:1.2.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
27	427	eco00620	1.2.1:2.3.1:3.6.1:6.2.1:6.2.1:2.3.3
27	428	eco00620	1.2.1:2.3.1:3.6.1:6.2.1:6.2.1:6.4.1
27	377	eco00620	3.1.2:1.1.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
28	195	eco00290	4.2.1:2.6.1:6.1.1
28	227	eco00400	2.7.1:2.5.1:4.2.3:5.4.99:5.4.99:2.6.1:6.1.1
28	192	eco00272	4.4.1:2.5.1:6.1.1
28	189	eco00272	4.4.1:6.1.1
28	228	eco00400	2.7.1:2.5.1:4.2.3:5.4.99:2.6.1:5.4.99:6.1.1
28	190	eco00272	2.8.1:2.6.1:6.1.1
29	48	eco00040	2.7.1:4.1.2
29	50	eco00051	3.1.3:2.7.1:2.7.1:4.1.2
29	43	eco00040	4.2.1:2.7.1:4.1.3
29	164	eco00260	3.1.3:2.7.8:4.1.1
29	45	eco00040	5.3.1:1.1.1:4.2.1:2.7.1:4.1.3
29	49	eco00051	3.1.3:2.7.1:4.1.2
29	31	eco00030	2.7.1:4.1.3:2.2.1
29	36	eco00030	2.7.1:4.1.2:2.2.1
29	65	eco00052	4.2.1:2.7.1:4.1.2:4.1.2
29	55	eco00051	9.9.9:1.1.1:2.7.1:4.1.2
29	51	eco00051	2.7.1:1.1.1:2.7.1:4.1.2
30	169	eco00260	2.7.1:4.2.3:4.1.2:6.1.1
30	171	eco00260	2.7.1:4.2.3:4.1.2:2.1.2:2.7.8:4.1.1

30	170	eco00260	2.7.1:4.2.3:4.1.2:2.1.2:6.1.1
30	173	eco00260	2.7.1:4.2.3:4.1.2:2.3.1
30	78	eco00130	5.4.4:2.2.1:4.2.99:4.2.1:6.2.1:4.1.3
30	68	eco00053	4.2.1:4.1.2
30	230	eco00400	2.5.1:4.2.3:4.2.1
30	452	eco00629	4.2.1:4.1.3
30	166	eco00260	2.7.1:4.2.3:6.1.1
30	33	eco00030	2.7.1:4.2.1:4.1.3:2.2.1
30	174	eco00260	2.7.1:4.2.3:4.1.2:1.4.4
30	226	eco00400	2.7.1:2.5.1:4.2.3:4.1.3
30	175	eco00260	2.7.1:4.2.3:1.1.1
30	222	eco00362	4.2.1:4.1.3
30	201	eco00300	2.7.2:4.2.1
30	451	eco00621	4.2.1:4.1.3
30	172	eco00260	2.7.1:4.2.3:4.1.2:2.1.2:4.3.1
30	168	eco00260	2.7.1:4.2.3:4.1.2:4.1.2
30	167	eco00260	2.7.1:4.2.3:4.3.1
31	93	eco00240	4.1.1:2.7.4:2.7.4:6.3.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:3.1.3:2.7.1:2.4.2
31	89	eco00240	4.1.1:2.7.4:2.7.4:6.3.4:3.5.4:3.6.1:2.4.2
31	90	eco00240	4.1.1:2.7.4:2.7.4:6.3.4:3.5.4:3.6.1:3.1.3:2.7.1:2.4.2
31	88	eco00240	4.1.1:2.7.4:2.7.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:3.1.3:2.4.2
31	95	eco00240	4.1.1:2.7.4:2.7.4:3.6.1:2.4.2
31	99	eco00240	4.1.1:3.1.3:2.7.1:2.7.4:2.7.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:2.4.2
31	96	eco00240	4.1.1:2.7.4:2.7.4:3.6.1:3.1.3:2.7.1:2.4.2
31	87	eco00240	4.1.1:2.7.4:2.7.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:3.1.3:2.7.1:2.4.2
31	86	eco00240	4.1.1:2.7.4:2.7.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:2.4.2
31	101	eco00240	4.1.1:3.1.3:2.7.1:2.7.4:2.7.4:6.3.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:2.4.2
31	103	eco00240	4.1.1:3.1.3:2.7.1:2.4.2
31	102	eco00240	4.1.1:3.1.3:2.7.1:2.7.4:2.7.4:3.6.1:2.4.2
31	92	eco00240	4.1.1:2.7.4:2.7.4:6.3.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:2.4.2
31	97	eco00240	4.1.1:2.7.4:2.7.4:3.6.1:3.1.3:2.4.2
31	91	eco00240	4.1.1:2.7.4:2.7.4:6.3.4:3.5.4:3.6.1:3.1.3:2.4.2
31	100	eco00240	4.1.1:3.1.3:2.7.1:2.7.4:2.7.4:6.3.4:3.5.4:3.6.1:2.4.2
31	104	eco00240	4.1.1:3.1.3:2.4.2
31	118	eco00240	3.1.3:3.5.4:2.4.2
31	215	eco00340	3.6.1:3.6.1:5.3.1:2.4.2
31	287	eco00530	4.1.3:3.1.3
31	94	eco00240	4.1.1:2.7.4:2.7.4:6.3.4:2.7.7:2.7.7:2.7.4:3.5.4:3.6.1:3.1.3:2.4.2
32	444	eco00620	1.1.1:1.2.1:1.2.2:6.2.1:6.2.1:2.3.1
32	336	eco00620	1.1.1:1.2.1:1.1.2:1.2.2:6.2.1:6.2.1:2.3.1
32	429	eco00620	1.2.1:2.3.1
32	337	eco00620	1.1.1:1.2.1:1.1.2:1.2.2:6.2.1:6.2.1:2.3.3

32	373	eco00620	3.1.2:1.1.1:1.2.2:6.2.1:6.2.1:2.3.3
32	318	eco00620	4.2.3:1.2.1:1.2.2:6.2.1:6.2.1:2.3.1
32	319	eco00620	4.2.3:1.2.1:1.2.2:6.2.1:6.2.1:2.3.3
32	430	eco00620	1.2.1:2.3.3
32	160	eco00260	1.1.99:1.2.1
32	205	eco00310	1.2.4:2.3.1
32	372	eco00620	3.1.2:1.1.1:1.2.2:6.2.1:6.2.1:2.3.1
32	445	eco00620	1.1.1:1.2.1:1.2.2:6.2.1:6.2.1:2.3.3
32	217	eco00360	1.4.3:1.2.1:6.2.1:2.3.1
32	355	eco00620	4.4.1:1.2.1:1.2.2:6.2.1:6.2.1:2.3.3
32	24	eco00020	1.8.1:1.2.4:1.2.4
32	450	eco00620	1.8.1:1.2.4:1.2.4
32	18	eco00010	1.8.1:1.2.4:1.2.4
32	354	eco00620	4.4.1:1.2.1:1.2.2:6.2.1:6.2.1:2.3.1
33	44	eco00040	4.2.1:1.1.1
33	196	eco00290	4.2.1:4.2.1:1.1.1:2.2.1:1.1.1
33	70	eco00061	2.3.1:2.3.1:1.1.1:4.2.1:1.3.1:2.3.1:1.1.1:4.2.1:1.3.1
33	212	eco00340	4.2.1:1.1.1:6.1.1
33	213	eco00340	4.2.1:1.1.1:1.1.1:6.1.1
33	73	eco00061	4.2.1:1.3.1
33	21	eco00020	1.1.1:2.3.3:4.2.1:4.2.1:1.1.1:1.1.1
33	46	eco00040	5.3.1:1.1.1:4.2.1:1.1.1
33	54	eco00051	2.7.7:4.2.1:1.1.1
34	414	eco00620	1.2.1:2.3.3:1.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
34	416	eco00620	1.2.1:2.3.3:1.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
34	425	eco00620	1.2.1:2.3.3:1.1.99:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
34	418	eco00620	1.2.1:2.3.3:1.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
34	197	eco00290	1.2.4:2.2.1:1.1.1
34	420	eco00620	1.2.1:2.3.3:1.1.99:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.1
34	424	eco00620	1.2.1:2.3.3:1.1.99:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.3
34	415	eco00620	1.2.1:2.3.3:1.1.1:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
34	421	eco00620	1.2.1:2.3.3:1.1.99:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:2.3.3
34	417	eco00620	1.2.1:2.3.3:1.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
34	423	eco00620	1.2.1:2.3.3:1.1.99:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:2.3.1
34	419	eco00620	1.2.1:2.3.3:1.1.1:4.1.1:2.7.1:1.2.2:2.7.2:3.6.1:6.2.1:6.2.1:6.4.1
34	422	eco00620	1.2.1:2.3.3:1.1.99:4.1.1:2.7.1:1.2.2:6.2.1:6.2.1:6.4.1
34	410	eco00620	1.2.1:2.3.3:1.1.1:1.2.2:6.2.1:6.2.1:6.4.1
34	409	eco00620	1.2.1:2.3.3:1.1.1:1.2.2:6.2.1:6.2.1:2.3.3
34	408	eco00620	1.2.1:2.3.3:1.1.1:1.2.2:6.2.1:6.2.1:2.3.1

Apéndice C

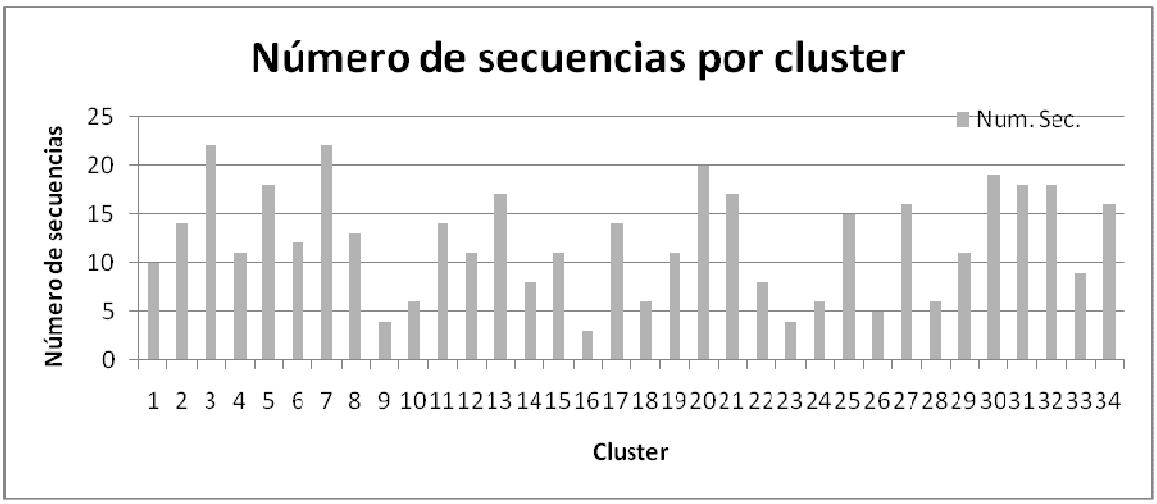
Matriz de alineamientos por pares de las 452 secuencias.



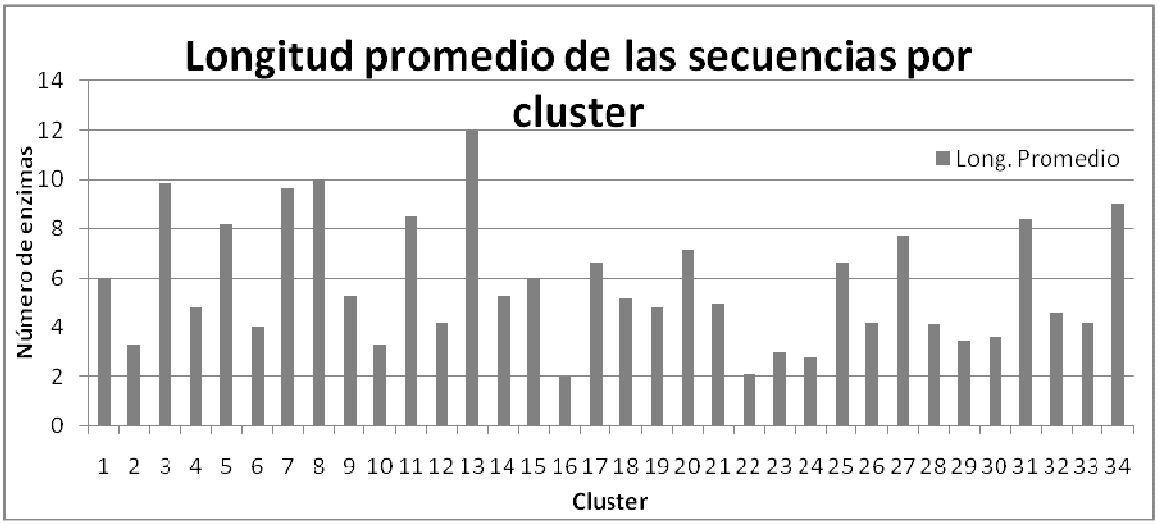
Apéndice D

Estadísticas de los *clusters*

En estas dos gráficas podemos observar la siguiente tendencia: a menor número de vías asociadas a los *clusters*, mayor será la longitud promedio de la secuencia.



Gráfica de la distribución de las secuencias por *cluster*.



Gráfica de la longitud promedio de las secuencias en cada *cluster*.

Apéndice E

Alineamientos múltiples de los 34 *clusters*

de secuencias: 10 Nombre: >Cluster1 Fitness: 0.268241

	0	1	2	3	4	5	6	7
150 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	1.4.1	1.4.1	4.1.1	2.6.1	1.2.1
147 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	1.4.1	2.6.1	4.1.1	2.6.1	1.2.1
133 - Glutamate metabolism	1.5.99	-.-.-	1.4.1	1.4.1	-.-.-	4.1.1	2.6.1	1.2.1
153 - Glutamate metabolism	1.5.99	6.3.1	-.-.-	1.4.1	-.-.-	4.1.1	2.6.1	1.2.1
136 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	1.4.1	4.1.1	2.6.1	1.2.1
122 - Glutamate metabolism	1.5.99	-.-.-	-.-.-	1.4.1	2.6.1	4.1.1	2.6.1	1.2.1
144 - Glutamate metabolism	1.5.99	-.-.-	-.-.-	-.-.-	-.-.-	4.1.1	2.6.1	1.2.1
125 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	1.4.1	4.1.1	2.6.1	1.2.1
083 - Urea cycle and metabolism of amino groups	3.5.1	-.-.-	-.-.-	-.-.-	-.-.-	4.1.1	2.6.1	1.2.1
080 - Urea cycle and metabolism of amino groups	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.6.1	1.2.1

***Cluster* 1**

de secuencias: 14 Nombre: >Cluster2 Fitness: 0.338746

	0	1	2	3	4
289 - Lipopolysaccharide biosynthesis	5.0.0	2.7.7	3.1.3	2.7.7	5.1.3
064 - Galactose metabolism	-.-.-	-.-.-	-.-.-	2.7.7	5.1.3
265 - Nucleotide sugars metabolism	-.-.-	-.-.-	-.-.-	2.7.7	5.1.3
038 - Pentose and glucuronate interconversions	-.-.-	-.-.-	-.-.-	2.7.1	5.1.3
269 - Aminosugars metabolism	-.-.-	-.-.-	3.2.1	2.7.1	5.1.3
032 - Pentose phosphate pathway	-.-.-	2.7.1	-.-.-	1.1.1	5.1.3
039 - Pentose and glucuronate interconversions	-.-.-	2.7.1	-.-.-	5.0.0	5.1.3
042 - Pentose and glucuronate interconversions	-.-.-	2.7.1	4.1.1	5.0.0	5.1.3
268 - Streptomycin biosynthesis	-.-.-	2.7.7	-.-.-	4.2.1	5.1.3
278 - Aminosugars metabolism	-.-.-	2.7.1	-.-.-	4.2.0	5.1.3
029 - Pentose phosphate pathway	-.-.-	2.7.1	-.-.-	5.3.1	5.1.3
067 - Ascorbate and aldarate metabolism	-.-.-	2.7.1	4.1.1	5.1.3	5.1.3
066 - Ascorbate and aldarate metabolism	2.7.1	3.1.1	4.1.1	5.1.3	5.1.3
266 - Nucleotide sugars metabolism	-.-.-	2.7.7	4.2.1	5.1.3	1.1.1

***Cluster* 2**

de secuencias: 22 Nombre: >Cluster3 Fitness: 0.118178

	0	1	2	3	4	5	6	7	8	9	10
324 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
325 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
432 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
433 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
326 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
396 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
397 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
360 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
361 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
342 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
434 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
306 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
343 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
307 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
378 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
379 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
398 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
362 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
344 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
308 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
380 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
022 - Citrate cycle (TCA cycle)	-.-.-	1.1.1	-.-.-	2.3.3	4.1.3	4.1.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-

***Cluster* 3**

de secuencias: 11 Nombre: >Cluster4 Fitness: 0.408397

	0	1	2	3	4	5	6	7	8	9
291 - Lipopolysaccharide biosynthesis	2.3.1	3.5.1	2.3.1	3.6.1	2.4.1	2.7.1	2.0.0	2.0.0	2.3.1	2.3.1
295 - Glycerolipid metabolism	-.-.-	3.2.1	-.-.-	-.-.-	-.-.-	2.7.1	-.-.-	-.-.-	2.3.1	2.3.1
063 - Galactose metabolism	-.-.-	3.2.1	-.-.-	-.-.-	-.-.-	2.7.1	-.-.-	-.-.-	-.-.-	2.7.7
059 - Galactose metabolism	-.-.-	3.2.1	-.-.-	3.2.1	-.-.-	2.7.1	-.-.-	-.-.-	-.-.-	2.7.7
028 - Pentose phosphate pathway	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.1	-.-.-	-.-.-	-.-.-	2.7.6
240 - Selenoamino acid metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.7	-.-.-	-.-.-	-.-.-	2.7.1
111 - Pyrimidine metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.4	-.-.-	-.-.-	2.7.4	2.7.7
299 - Glycerophospholipid metabolism	-.-.-	3.1.4	2.3.1	-.-.-	2.3.1	2.7.7	-.-.-	-.-.-	2.7.8	3.1.1
300 - Glycerophospholipid metabolism	-.-.-	3.1.4	2.3.1	-.-.-	2.3.1	2.7.7	-.-.-	2.7.8	3.1.3	2.7.8
297 - Glycerophospholipid metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.1	-.-.-	2.7.7	2.7.8	3.1.1
076 - Biosynthesis of steroids	2.2.1	1.1.1	2.7.7	-.-.-	-.-.-	2.7.1	4.6.1	1.17.7	1.17.1	2.5.1

***Cluster* 4**

de secuencias: 18 Nombre: >Cluster5 Fitness: 0.217126

	0	1	2	3	4	5	6	7	8	9
280 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	5.1.3	1.1.1
281 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	5.1.3	3.1.3
282 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	5.1.3	2.7.1
279 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	2.5.1	1.1.1
271 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	5.1.3	1.1.1
284 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	5.1.3	1.1.1
272 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	5.1.3	3.1.3
285 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	5.1.3	3.1.3
273 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	5.1.3	2.7.1
286 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	5.1.3	2.7.1
270 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	3.5.99	2.6.1	5.4.2	2.7.7	2.7.7	2.5.1	1.1.1
283 - Aminosugars metabolism	2.7.1	4.2.0	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	2.5.1	1.1.1
275 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	5.1.3	1.1.1
276 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	5.1.3	3.1.3
277 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	5.1.3	2.7.1
274 - Aminosugars metabolism	3.2.1	2.7.1	3.5.1	-.-.-	-.-.-	5.4.2	2.7.7	2.7.7	2.5.1	1.1.1
003 - Glycolysis / Gluconeogenesis	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	5.4.2	2.7.2	-.-.-	-.-.-	-.-.-
026 - Pentose phosphate pathway	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	5.4.2	2.7.6	-.-.-	-.-.-	-.-.-

***Cluster* 5**

de secuencias: 12 Nombre: >Cluster6 Fitness: 0.291647

	0	1	2	3	4	5
148 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	1.4.1	1.4.1	5.1.1
149 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	1.4.1	1.4.1	6.1.1
145 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	1.4.1	2.6.1	5.1.1
151 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	-.-.-	-.-.-	5.1.1
146 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	1.4.1	2.6.1	6.1.1
152 - Glutamate metabolism	1.5.99	6.3.1	1.4.1	-.-.-	-.-.-	6.1.1
142 - Glutamate metabolism	1.5.99	-.-.-	-.-.-	-.-.-	-.-.-	5.1.1
155 - Glutamate metabolism	1.5.99	6.3.1	-.-.-	-.-.-	-.-.-	2.4.2
158 - Glutamate metabolism	1.5.99	6.3.1	-.-.-	-.-.-	-.-.-	6.1.1
143 - Glutamate metabolism	1.5.99	-.-.-	-.-.-	-.-.-	-.-.-	6.1.1
157 - Glutamate metabolism	1.5.99	6.3.1	-.-.-	-.-.-	-.-.-	3.5.1
156 - Glutamate metabolism	1.5.99	6.3.1	-.-.-	-.-.-	-.-.-	6.3.5

***Cluster* 6**

de secuencias: 22 Nombre: >Cluster7 Fitness: 0.154683

	0	1	2	3	4	5	6	7	8	9	10	11
363 - Pyruvate metabolism	3.1.2	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
369 - Pyruvate metabolism	3.1.2	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
370 - Pyruvate metabolism	3.1.2	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
371 - Pyruvate metabolism	3.1.2	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
441 - Pyruvate metabolism	1.1.1	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
442 - Pyruvate metabolism	1.1.1	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
351 - Pyruvate metabolism	4.4.1	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
315 - Pyruvate metabolism	4.2.3	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
387 - Pyruvate metabolism	-.-.-	2.3.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
333 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	-.-.-	2.7.9	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
316 - Pyruvate metabolism	4.2.3	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
352 - Pyruvate metabolism	4.4.1	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
388 - Pyruvate metabolism	-.-.-	2.3.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
334 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	-.-.-	2.7.9	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
393 - Pyruvate metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
394 - Pyruvate metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
443 - Pyruvate metabolism	1.1.1	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
389 - Pyruvate metabolism	-.-.-	2.3.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
317 - Pyruvate metabolism	4.2.3	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
353 - Pyruvate metabolism	4.4.1	1.2.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
335 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	-.-.-	2.7.9	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
395 - Pyruvate metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1

Cluster 7

de secuencias: 13 Nombre: >Cluster8 Fitness: 0.174491

	0	1	2	3	4	5	6	7	8	9	10	11	12
400 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
399 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
401 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
406 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
405 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
412 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	-.-.-	-.-.-	-.-.-	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
407 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
411 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	-.-.-	-.-.-	-.-.-	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
403 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.3
402 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.1
413 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	-.-.-	-.-.-	-.-.-	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
404 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	6.4.1
385 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	-.-.-	-.-.-	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.3

Cluster 8

de secuencias: 4 Nombre: >Cluster9 Fitness: 0.344485

	0	1	2	3	4	5	6
219 - Phenylalanine metabolism	1.14.12	1.3.1	1.13.11	3.7.1	4.2.1	4.1.3	-.-.-
216 - Phenylalanine metabolism	1.14.13	-.-.-	1.13.11	3.7.1	4.2.1	4.1.3	-.-.-
221 - Phenylalanine metabolism	1.14.12	1.3.1	1.13.11	3.7.1	-.-.-	-.-.-	-.-.-
035 - Pentose phosphate pathway	5.3.1	-.-.-	1.1.1	3.1.1	4.2.1	4.1.3	2.2.1

Cluster 9

de secuencias: 6 Nombre: >Cluster10 Fitness: 0.404181

	0	1	2	3	4	5	6
259 - Starch and sucrose metabolism	2.4.1	2.7.7	2.4.1	2.4.1	2.4.1	2.7.7	1.1.1
252 - Starch and sucrose metabolism	-.-.-	-.-.-	-.-.-	-.-.-	2.4.1	2.7.7	1.1.1
041 - Pentose and glucuronate interconversions	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.7	1.1.1
161 - Glycine, serine and threonine metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.7.1	1.1.1
198 - Valine, leucine and isoleucine biosynthesis	-.-.-	-.-.-	-.-.-	-.-.-	2.2.1	2.2.1	1.1.1
199 - Valine, leucine and isoleucine biosynthesis	-.-.-	-.-.-	-.-.-	-.-.-	4.3.1	2.2.1	1.1.1

***Cluster* 10**

de secuencias: 14 Nombre: >Cluster11 Fitness: 0.286861

	0	1	2	3	4	5	6	7	8	9	10	11	12
106 - Pyrimidine metabolism	2.7.4	2.7.4	3.5.4	3.6.1	2.7.4	2.7.4	3.6.1	2.1.1	3.1.3	2.7.1	2.7.4	2.7.4	2.7.7
105 - Pyrimidine metabolism	2.7.4	2.7.4	3.5.4	3.6.1	2.7.4	2.7.4	3.6.1	2.1.1	-.-.-	-.-.-	2.7.4	2.7.4	2.7.7
109 - Pyrimidine metabolism	2.7.4	2.7.4	3.5.4	3.6.1	-.-.-	2.1.1	3.1.3	2.7.1	-.-.-	-.-.-	2.7.4	2.7.4	2.7.7
107 - Pyrimidine metabolism	2.7.4	2.7.4	3.5.4	3.6.1	2.7.4	2.7.4	3.6.1	2.1.1	-.-.-	-.-.-	3.1.3	2.4.2	-.-.-
113 - Pyrimidine metabolism	3.1.3	-.-.-	3.5.4	2.7.1	2.7.4	2.7.4	3.6.1	2.1.1	3.1.3	2.7.1	2.7.4	2.7.4	2.7.7
108 - Pyrimidine metabolism	2.7.4	2.7.4	3.5.4	3.6.1	-.-.-	2.1.1	-.-.-	2.7.4	-.-.-	-.-.-	2.7.4	2.7.7	-.-.-
116 - Pyrimidine metabolism	3.1.3	-.-.-	3.5.4	2.7.1	-.-.-	2.1.1	3.1.3	2.7.1	-.-.-	-.-.-	2.7.4	2.7.4	2.7.7
110 - Pyrimidine metabolism	2.7.4	2.7.4	3.5.4	3.6.1	-.-.-	-.-.-	-.-.-	2.1.1	-.-.-	-.-.-	3.1.3	2.4.2	-.-.-
112 - Pyrimidine metabolism	3.1.3	-.-.-	3.5.4	2.7.1	2.7.4	2.7.4	3.6.1	2.1.1	-.-.-	-.-.-	2.7.4	2.7.4	2.7.7
115 - Pyrimidine metabolism	3.1.3	-.-.-	3.5.4	2.7.1	-.-.-	2.1.1	-.-.-	2.7.4	-.-.-	-.-.-	2.7.4	2.7.7	-.-.-
114 - Pyrimidine metabolism	3.1.3	-.-.-	3.5.4	2.7.1	2.7.4	2.7.4	3.6.1	2.1.1	-.-.-	-.-.-	3.1.3	2.4.2	-.-.-
298 - Glycerophospholipid metabolism	2.7.1	-.-.-	-.-.-	2.7.7	-.-.-	-.-.-	-.-.-	2.7.8	-.-.-	-.-.-	3.1.3	2.7.8	-.-.-
117 - Pyrimidine metabolism	3.1.3	-.-.-	3.5.4	2.7.1	-.-.-	-.-.-	-.-.-	2.1.1	-.-.-	-.-.-	3.1.3	2.4.2	-.-.-
290 - Lipopolysaccharide biosynthesis	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.5.1	-.-.-	-.-.-	3.1.3	2.7.7	-.-.-

***Cluster* 11**

de secuencias: 11 Nombre: >Cluster12 Fitness: 0.273764

	0	1	2	3	4	5
123 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	1.4.1	5.1.1
120 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	-.-.-	-.-.-	5.1.1
127 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	-.-.-	2.4.2
130 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	-.-.-	6.1.1
121 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	-.-.-	-.-.-	6.1.1
126 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	-.-.-	2.6.1
129 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	-.-.-	3.5.1
128 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	-.-.-	6.3.5
218 - Phenylalanine metabolism	-.-.-	1.4.99	2.6.1	-.-.-	-.-.-	-.-.-
154 - Glutamate metabolism	1.5.99	6.3.1	2.6.1	-.-.-	-.-.-	-.-.-
208 - Arginine and proline metabolism	-.-.-	1.5.99	2.6.1	-.-.-	-.-.-	4.1.3

***Cluster* 12**

de secuencias: 17 Nombre: >Cluster13 Fitness: 0.081156

	0	1	2	3	4	5	6	7	8	9	10	11	12
327 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
328 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
435 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
436 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
329 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
364 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
345 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
437 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
309 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
346 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
310 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
381 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
382 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
365 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
347 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
311 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
383 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	4.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1

***Cluster* 13**

de secuencias: 8 Nombre: >Cluster14 Fitness: 0.264229

	0	1	2	3	4	5	6
016 - Glycolysis / Gluconeogenesis	3.1.3	2.7.1	5.3.1	5.3.1	2.7.1	4.1.2	1.2.1
006 - Glycolysis / Gluconeogenesis	-.-.-	2.7.1	5.3.1	5.3.1	2.7.1	4.1.2	1.2.1
004 - Glycolysis / Gluconeogenesis	-.-.-	2.7.1	5.3.1	-.-.-	2.7.1	4.1.2	1.2.1
019 - Glycolysis / Gluconeogenesis	-.-.-	2.7.1	3.2.1	5.3.1	2.7.1	4.1.2	1.2.1
008 - Glycolysis / Gluconeogenesis	-.-.-	5.4.2	5.3.1	-.-.-	2.7.1	4.1.2	1.2.1
010 - Glycolysis / Gluconeogenesis	-.-.-	5.4.2	5.3.1	5.3.1	2.7.1	4.1.2	1.2.1
224 - Phenylalanine, tyrosine and tryptophan biosynthesis	-.-.-	5.3.1	5.3.1	-.-.-	-.-.-	4.2.1	-.-.-
225 - Phenylalanine, tyrosine and tryptophan biosynthesis	-.-.-	5.3.1	5.3.1	-.-.-	-.-.-	4.2.1	4.2.1

***Cluster* 14**

de secuencias: 11 Nombre: >Cluster15 Fitness: 0.303562

	0	1	2	3	4	5	6	7	8
332 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
338 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	-.-.-	-.-.-	1.2.2	6.2.1	6.2.1	6.4.1
446 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	-.-.-	-.-.-	1.2.2	6.2.1	6.2.1	6.4.1
449 - Pyruvate metabolism	1.1.1	1.2.1	1.2.2	2.7.2	-.-.-	3.6.1	6.2.1	6.2.1	6.4.1
341 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
392 - Pyruvate metabolism	2.3.1	1.2.2	-.-.-	-.-.-	-.-.-	-.-.-	6.2.1	6.2.1	6.4.1
374 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	-.-.-	-.-.-	1.2.2	6.2.1	6.2.1	6.4.1
356 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	-.-.-	-.-.-	1.2.2	6.2.1	6.2.1	6.4.1
320 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	-.-.-	-.-.-	1.2.2	6.2.1	6.2.1	6.4.1
431 - Pyruvate metabolism	-.-.-	1.2.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	6.4.1
072 - Fatty acid biosynthesis	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	6.4.1	6.4.1

***Cluster* 15**

de secuencias: 3 Nombre: >Cluster16 Fitness: 0.147073

	0	1
211 - Histidine metabolism	3.4.13	6.1.1
163 - Glycine, serine and threonine metabolism	3.1.3	6.1.1
235 - beta-Alanine metabolism	3.4.13	6.3.2

***Cluster* 16**

de secuencias: 14 Nombre: >Cluster17 Fitness: 0.347314

	0	1	2	3	4	5	6	7	8	9	10	11	12
260 - Starch and sucrose metabolism	2.4.1	2.7.7	2.4.1	2.4.1	2.4.1	2.7.7	2.4.1	3.1.3	3.2.1	2.4.1	3.2.1	2.7.1	5.3.1
261 - Starch and sucrose metabolism	2.4.1	2.7.7	2.4.1	2.4.1	2.4.1	2.7.7	2.4.1	3.1.3	3.2.1	2.4.1	-.-.-	2.7.1	-.-.-
253 - Starch and sucrose metabolism	-.-.-	-.-.-	2.4.1	-.-.-	-.-.-	2.7.7	2.4.1	3.1.3	3.2.1	2.4.1	3.2.1	2.7.1	5.3.1
262 - Starch and sucrose metabolism	2.4.1	2.7.7	2.4.1	2.4.1	2.4.1	2.7.7	2.4.1	3.1.3	-.-.-	2.7.1	3.2.1	-.-.-	-.-.-
258 - Starch and sucrose metabolism	2.4.1	2.7.7	2.4.1	2.4.1	2.4.1	2.7.7	2.4.1	3.2.1	3.2.1	-.-.-	-.-.-	-.-.-	-.-.-
257 - Starch and sucrose metabolism	2.4.1	2.7.7	2.4.1	2.4.1	2.4.1	2.7.7	2.4.1	3.2.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
254 - Starch and sucrose metabolism	-.-.-	-.-.-	2.4.1	-.-.-	-.-.-	2.7.7	2.4.1	3.1.3	3.2.1	2.4.1	-.-.-	2.7.1	-.-.-
255 - Starch and sucrose metabolism	-.-.-	-.-.-	2.4.1	-.-.-	-.-.-	2.7.7	2.4.1	3.1.3	-.-.-	2.7.1	3.2.1	-.-.-	-.-.-
256 - Starch and sucrose metabolism	-.-.-	-.-.-	2.4.1	-.-.-	-.-.-	2.7.7	2.4.1	-.-.-	-.-.-	2.4.1	3.2.1	-.-.-	-.-.-
251 - Starch and sucrose metabolism	-.-.-	-.-.-	2.4.1	-.-.-	-.-.-	2.7.7	2.4.1	3.2.1	-.-.-	-.-.-	3.2.1	-.-.-	-.-.-
250 - Starch and sucrose metabolism	-.-.-	-.-.-	2.4.1	-.-.-	-.-.-	2.7.7	2.4.1	3.2.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
062 - Galactose metabolism	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	3.2.1	-.-.-	-.-.-	3.2.1	-.-.-	-.-.-
304 - Glycosphingolipid biosynthesis - globo series	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	3.2.1	-.-.-	-.-.-	3.2.1	-.-.-	-.-.-
305 - Glycosphingolipid biosynthesis - ganglio series	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	3.2.1	-.-.-	-.-.-	3.2.1	-.-.-	-.-.-

***Cluster* 17**

de secuencias: 6 Nombre: >Cluster18 Fitness: 0.431943

	0	1	2	3	4	5	6	7	8
081 - Urea cycle and metabolism of amino groups	3.5.1	2.1.3	6.3.4	4.3.2	4.1.1	3.5.3	2.6.1	1.2.1	-.-.-
082 - Urea cycle and metabolism of amino groups	3.5.1	2.1.3	6.3.4	4.3.2	4.1.1	3.5.3	2.3.1	-.-.-	-.-.-
206 - Arginine and proline metabolism	-.-.-	2.1.3	6.3.4	4.3.2	2.3.1	3.5.3	2.6.1	1.2.1	3.5.1
207 - Arginine and proline metabolism	-.-.-	2.1.3	6.3.4	4.3.2	6.1.1	-.-.-	-.-.-	-.-.-	-.-.-
159 - Alanine and aspartate metabolism	-.-.-	-.-.-	-.-.-	3.4.13	2.6.1	-.-.-	-.-.-	-.-.-	-.-.-
234 - beta-Alanine metabolism	-.-.-	-.-.-	-.-.-	3.4.13	2.6.1	-.-.-	-.-.-	-.-.-	-.-.-

***Cluster* 18**

de secuencias: 11 Nombre: >Cluster19 Fitness: 0.23939

	0	1	2	3	4	5
134 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	1.4.1	5.1.1
135 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	1.4.1	6.1.1
131 - Glutamate metabolism	1.5.99	-.-.-	1.4.1	-.-.-	1.4.1	5.1.1
138 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	2.4.2	-.-.-
141 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	-.-.-	6.1.1
132 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	-.-.-	-.-.-	6.1.1
140 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	-.-.-	3.5.1
137 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	-.-.-	2.6.1
124 - Glutamate metabolism	1.5.99	1.4.1	2.6.1	6.3.1	1.4.1	6.1.1
139 - Glutamate metabolism	1.5.99	1.4.1	1.4.1	6.3.1	-.-.-	6.3.5
210 - Arginine and proline metabolism	1.5.1	-.-.-	1.5.99	-.-.-	-.-.-	-.-.-

***Cluster* 19**

de secuencias: 20 Nombre: >Cluster20 Fitness: 0.191815

	0	1	2	3	4	5	6	7	8
330 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
331 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
438 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
439 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
366 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
367 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
348 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
440 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
312 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
349 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
313 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.3
384 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	2.3.1
390 - Pyruvate metabolism	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1	1.2.2	6.2.1	6.2.1	2.3.1
368 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
391 - Pyruvate metabolism	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1	1.2.2	6.2.1	6.2.1	2.3.3
350 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
314 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
386 - Pyruvate metabolism	-.-.-	2.3.1	-.-.-	2.7.9	2.7.1	1.2.2	6.2.1	6.2.1	6.4.1
162 - Glycine, serine and threonine metabolism	-.-.-	-.-.-	-.-.-	-.-.-	2.7.2	1.2.1	-.-.-	-.-.-	-.-.-
200 - Lysine biosynthesis	-.-.-	-.-.-	-.-.-	-.-.-	2.7.2	1.2.1	4.2.1	-.-.-	-.-.-

***Cluster* 20**

de secuencias: 17 Nombre: >Cluster21 Fitness: 0.265697

	0	1	2	3	4	5	6
013 - Glycolysis / Gluconeogenesis	3.1.3	5.1.3	2.7.1	5.3.1	2.7.1	4.1.2	5.3.1
015 - Glycolysis / Gluconeogenesis	3.1.3	-.-.-	2.7.1	5.3.1	2.7.1	4.1.2	5.3.1
012 - Glycolysis / Gluconeogenesis	3.1.3	5.1.3	2.7.1	5.3.1	2.7.1	4.1.2	1.2.1
005 - Glycolysis / Gluconeogenesis	-.-.-	-.-.-	2.7.1	5.3.1	2.7.1	4.1.2	5.3.1
017 - Glycolysis / Gluconeogenesis	3.1.3	2.7.1	5.3.1	5.3.1	2.7.1	4.1.2	5.3.1
014 - Glycolysis / Gluconeogenesis	3.1.3	-.-.-	2.7.1	5.3.1	2.7.1	4.1.2	1.2.1
009 - Glycolysis / Gluconeogenesis	-.-.-	-.-.-	5.4.2	5.3.1	2.7.1	4.1.2	5.3.1
007 - Glycolysis / Gluconeogenesis	-.-.-	2.7.1	5.3.1	5.3.1	2.7.1	4.1.2	5.3.1
020 - Glycolysis / Gluconeogenesis	-.-.-	2.7.1	3.2.1	5.3.1	2.7.1	4.1.2	5.3.1
011 - Glycolysis / Gluconeogenesis	-.-.-	5.4.2	5.3.1	5.3.1	2.7.1	4.1.2	5.3.1
056 - Fructose and mannose metabolism	-.-.-	-.-.-	-.-.-	5.3.1	2.7.1	4.1.2	-.-.-
060 - Galactose metabolism	-.-.-	-.-.-	2.7.1	5.3.1	2.7.1	-.-.-	-.-.-
057 - Fructose and mannose metabolism	-.-.-	-.-.-	5.3.1	2.7.1	2.7.1	4.1.2	-.-.-
052 - Fructose and mannose metabolism	-.-.-	-.-.-	5.4.2	5.3.1	2.7.1	4.1.2	-.-.-
030 - Pentose phosphate pathway	-.-.-	-.-.-	-.-.-	5.3.1	2.7.1	4.1.2	2.2.1
040 - Pentose and glucuronate interconversions	-.-.-	-.-.-	-.-.-	5.3.1	2.7.1	-.-.-	-.-.-
037 - Pentose and glucuronate interconversions	-.-.-	-.-.-	-.-.-	5.3.1	2.7.1	5.1.3	-.-.-

***Cluster* 21**

de secuencias: 8 Nombre: >Cluster22 Fitness: 0.242255

	0	1	2
084 - Urea cycle and metabolism of amino groups	3.5.1	4.1.1	2.3.1
237 - Taurine and hypotaurine metabolism	-.-.-	4.1.1	2.3.2
001 - Glycolysis / Gluconeogenesis	-.-.-	4.1.1	2.7.1
085 - Urea cycle and metabolism of amino groups	-.-.-	4.1.1	2.5.1
232 - beta-Alanine metabolism	-.-.-	4.1.1	2.6.1
098 - Pyrimidine metabolism	-.-.-	4.1.1	2.4.2
288 - Aminosugars metabolism	-.-.-	4.1.3	2.7.1
002 - Glycolysis / Gluconeogenesis	-.-.-	4.2.1	2.7.1

***Cluster* 22**

de secuencias: 4 Nombre: >Cluster23 Fitness: 0.423588

	0	1	2	3	4
034 - Pentose phosphate pathway	5.3.1	1.1.1	3.1.1	1.1.1	5.1.3
301 - Glycerophospholipid metabolism	-.-.-	-.-.-	3.1.4	1.1.1	-.-.-
302 - Glycerophospholipid metabolism	-.-.-	-.-.-	3.1.4	1.1.5	-.-.-
296 - Glycerolipid metabolism	-.-.-	-.-.-	3.2.1	1.1.1	2.7.1

***Cluster* 23**

de secuencias: 6 Nombre: >Cluster24 Fitness: 0.339734

	0	1	2	3
203 - Lysine biosynthesis	3.5.1	5.1.1	6.3.2	6.3.2
242 - D-Glutamine and D-glutamate metabolism	3.5.1	5.1.1	-.-.-	6.3.2
243 - D-Alanine metabolism	-.-.-	5.1.1	-.-.-	6.3.2
241 - D-Glutamine and D-glutamate metabolism	3.5.1	-.-.-	-.-.-	6.3.2
233 - beta-Alanine metabolism	-.-.-	4.1.1	-.-.-	6.3.2
202 - Lysine biosynthesis	3.5.1	5.1.1	4.1.1	6.1.1

***Cluster* 24**

de secuencias: 15 Nombre: >Cluster25 Fitness: 0.333104

	0	1	2	3	4	5	6	7	8	9	10	11
184 - Methionine metabolism	2.3.1	2.5.1	2.5.1	2.5.1	2.1.1	2.5.1	2.1.1	3.2.2	4.4.1	2.1.1	6.1.1	2.1.2
188 - Methionine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	2.1.1	2.5.1	2.1.1	3.2.2	4.4.1	2.1.1	6.1.1	2.1.2
181 - Methionine metabolism	2.3.1	2.5.1	-.-.-	4.4.1	2.1.1	2.5.1	2.1.1	3.2.2	4.4.1	2.1.1	6.1.1	2.1.2
178 - Methionine metabolism	2.6.1	2.5.1	-.-.-	-.-.-	2.1.1	-.-.-	-.-.-	3.2.2	4.4.1	2.1.1	6.1.1	2.1.2
182 - Methionine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	2.5.1	2.5.1	-.-.-	-.-.-	-.-.-	2.1.1	6.1.1	2.1.2
183 - Methionine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	2.5.1	2.5.1	2.1.1	2.5.1	4.1.1	2.5.1	-.-.-	3.2.2
179 - Methionine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	4.4.1	2.1.1	6.1.1	2.1.2
187 - Methionine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	2.1.1	2.5.1	-.-.-	-.-.-	4.1.1	2.5.1	-.-.-	3.2.2
186 - Methionine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.1.1	6.1.1	2.1.2
180 - Methionine metabolism	2.3.1	2.5.1	-.-.-	4.4.1	2.1.1	2.5.1	-.-.-	-.-.-	4.1.1	2.5.1	-.-.-	3.2.2
177 - Methionine metabolism	2.6.1	2.5.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	4.1.1	2.5.1	-.-.-	3.2.2
239 - Selenoamino acid metabolism	2.5.1	2.5.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	4.4.1	-.-.-	-.-.-	-.-.-
191 - Cysteine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	6.1.1	-.-.-
079 - Ubiquinone and menaquinone biosynthesis	-.-.-	2.5.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.1.1	-.-.-	-.-.-
185 - Methionine metabolism	2.3.1	2.5.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-

***Cluster* 25**

de secuencias: 5 Nombre: >Cluster26 Fitness: 0.398832

	0	1	2	3	4	5	6
247 - Glutathione metabolism	2.5.1	2.3.2	6.3.2	6.3.2	2.3.2	3.4.11	-.-.-
248 - Glutathione metabolism	2.5.1	2.3.2	6.3.2	6.3.2	2.3.2	3.4.13	-.-.-
244 - Glutathione metabolism	-.-.-	-.-.-	-.-.-	-.-.-	2.3.2	3.4.11	-.-.-
245 - Glutathione metabolism	-.-.-	-.-.-	-.-.-	-.-.-	2.3.2	3.4.13	-.-.-
292 - Peptidoglycan biosynthesis	-.-.-	-.-.-	6.3.2	6.3.2	2.7.8	2.4.1	6.3.1

***Cluster* 26**

de secuencias: 16 Nombre: >Cluster27 Fitness: 0.134784

	0	1	2	3	4	5	6	7	8
339 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
340 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
447 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
448 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
375 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
376 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
357 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
321 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
358 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
322 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
426 - Pyruvate metabolism	-.-.-	1.2.1	-.-.-	-.-.-	2.3.1	3.6.1	6.2.1	6.2.1	2.3.1
427 - Pyruvate metabolism	-.-.-	1.2.1	-.-.-	-.-.-	2.3.1	3.6.1	6.2.1	6.2.1	2.3.3
377 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
359 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
323 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
428 - Pyruvate metabolism	-.-.-	1.2.1	-.-.-	-.-.-	2.3.1	3.6.1	6.2.1	6.2.1	6.4.1

***Cluster* 27**

de secuencias: 6 Nombre: >Cluster28 Fitness: 0.407752

	0	1	2	3	4	5	6
227 - Phenylalanine, tyrosine and tryptophan biosynthesis	2.7.1	2.5.1	4.2.3	5.4.99	5.4.99	2.6.1	6.1.1
228 - Phenylalanine, tyrosine and tryptophan biosynthesis	2.7.1	2.5.1	4.2.3	5.4.99	2.6.1	5.4.99	6.1.1
195 - Valine, leucine and isoleucine biosynthesis	-.-.-	-.-.-	4.2.1	-.-.-	2.6.1	-.-.-	6.1.1
190 - Cysteine metabolism	-.-.-	-.-.-	2.8.1	-.-.-	2.6.1	-.-.-	6.1.1
192 - Cysteine metabolism	-.-.-	-.-.-	4.4.1	-.-.-	2.5.1	-.-.-	6.1.1
189 - Cysteine metabolism	-.-.-	-.-.-	4.4.1	-.-.-	-.-.-	-.-.-	6.1.1

***Cluster* 28**

de secuencias: 11 Nombre: >Cluster29 Fitness: 0.330912

	0	1	2	3	4	5
045 - Pentose and glucuronate interconversions	5.3.1	1.1.1	4.2.1	2.7.1	4.1.3	-.-.-
043 - Pentose and glucuronate interconversions	-.-.-	-.-.-	4.2.1	2.7.1	4.1.3	-.-.-
048 - Pentose and glucuronate interconversions	-.-.-	-.-.-	-.-.-	2.7.1	4.1.2	-.-.-
065 - Galactose metabolism	-.-.-	-.-.-	4.2.1	2.7.1	4.1.2	4.1.2
051 - Fructose and mannose metabolism	2.7.1	1.1.1	-.-.-	2.7.1	4.1.2	-.-.-
055 - Fructose and mannose metabolism	9.9.9	1.1.1	-.-.-	2.7.1	4.1.2	-.-.-
049 - Fructose and mannose metabolism	-.-.-	3.1.3	-.-.-	2.7.1	4.1.2	-.-.-
164 - Glycine, serine and threonine metabolism	-.-.-	3.1.3	-.-.-	2.7.8	4.1.1	-.-.-
031 - Pentose phosphate pathway	-.-.-	-.-.-	-.-.-	2.7.1	4.1.3	2.2.1
036 - Pentose phosphate pathway	-.-.-	-.-.-	-.-.-	2.7.1	4.1.2	2.2.1
050 - Fructose and mannose metabolism	-.-.-	3.1.3	2.7.1	2.7.1	4.1.2	-.-.-

***Cluster* 29**

de secuencias: 19 Nombre: >Cluster30 Fitness: 0.310797

	0	1	2	3	4	5
171 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	4.1.2	2.1.2	2.7.8	4.1.1
172 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	4.1.2	2.1.2	-.-.-	4.3.1
170 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	4.1.2	2.1.2	-.-.-	6.1.1
168 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	4.1.2	-.-.-	-.-.-	4.1.2
173 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	4.1.2	-.-.-	-.-.-	2.3.1
169 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	4.1.2	-.-.-	-.-.-	6.1.1
033 - Pentose phosphate pathway	2.7.1	4.2.1	4.1.3	-.-.-	-.-.-	2.2.1
167 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	-.-.-	-.-.-	-.-.-	4.3.1
174 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	4.1.2	-.-.-	-.-.-	1.4.4
068 - Ascorbate and aldarate metabolism	-.-.-	4.2.1	-.-.-	-.-.-	-.-.-	4.1.2
166 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	-.-.-	-.-.-	-.-.-	6.1.1
175 - Glycine, serine and threonine metabolism	2.7.1	4.2.3	-.-.-	-.-.-	-.-.-	1.1.1
201 - Lysine biosynthesis	2.7.2	4.2.1	-.-.-	-.-.-	-.-.-	-.-.-
222 - Benzoate degradation via hydroxylation	-.-.-	4.2.1	-.-.-	-.-.-	-.-.-	4.1.3
451 - Biphenyl degradation	-.-.-	4.2.1	-.-.-	-.-.-	-.-.-	4.1.3
452 - Carbazole degradation	-.-.-	4.2.1	-.-.-	-.-.-	-.-.-	4.1.3
226 - Phenylalanine, tyrosine and tryptophan biosynthesis	2.7.1	2.5.1	4.2.3	-.-.-	-.-.-	4.1.3
230 - Phenylalanine, tyrosine and tryptophan biosynthesis	2.5.1	4.2.3	-.-.-	-.-.-	-.-.-	4.2.1
078 - Ubiquinone and menaquinone biosynthesis	5.4.4	2.2.1	4.2.99	4.2.1	6.2.1	4.1.3

***Cluster* 30**

de secuencias: 18 Nombre: >Cluster31 Fitness: 0.254976

	0	1	2	3	4	5	6	7	8	9	10	11
093 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	6.3.4	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	3.1.3	2.7.1	2.4.2
087 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	-.-.-	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	3.1.3	2.7.1	2.4.2
094 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	6.3.4	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	3.1.3	-.-.-	2.4.2
092 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	6.3.4	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	-.-.-	-.-.-	2.4.2
088 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	-.-.-	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	3.1.3	-.-.-	2.4.2
090 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	6.3.4	-.-.-	-.-.-	-.-.-	3.5.4	3.6.1	3.1.3	2.7.1	2.4.2
086 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	-.-.-	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	-.-.-	-.-.-	2.4.2
091 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	6.3.4	-.-.-	-.-.-	-.-.-	3.5.4	3.6.1	3.1.3	-.-.-	2.4.2
101 - Pyrimidine metabolism	4.1.1	3.1.3	2.7.1	2.7.4	2.7.4	6.3.4	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	2.4.2
096 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	3.6.1	3.1.3	2.7.1	2.4.2
089 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	6.3.4	-.-.-	-.-.-	-.-.-	3.5.4	3.6.1	-.-.-	-.-.-	2.4.2
099 - Pyrimidine metabolism	4.1.1	3.1.3	2.7.1	2.7.4	2.7.4	2.7.7	2.7.7	2.7.4	3.5.4	3.6.1	-.-.-	2.4.2
097 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	3.6.1	3.1.3	-.-.-	2.4.2
095 - Pyrimidine metabolism	4.1.1	2.7.4	2.7.4	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	3.6.1	-.-.-	-.-.-	2.4.2
103 - Pyrimidine metabolism	4.1.1	3.1.3	2.7.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.4.2
100 - Pyrimidine metabolism	4.1.1	3.1.3	2.7.1	2.7.4	2.7.4	6.3.4	-.-.-	3.5.4	3.6.1	-.-.-	-.-.-	2.4.2
102 - Pyrimidine metabolism	4.1.1	3.1.3	2.7.1	-.-.-	2.7.4	2.7.4	-.-.-	-.-.-	3.6.1	-.-.-	-.-.-	2.4.2
104 - Pyrimidine metabolism	4.1.1	3.1.3	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	2.4.2

***Cluster* 31**

de secuencias: 18 Nombre: >Cluster32 Fitness: 0.273599

	0	1	2	3	4	5	6
336 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	1.2.2	6.2.1	6.2.1	2.3.1
337 - Pyruvate metabolism	1.1.1	1.2.1	1.1.2	1.2.2	6.2.1	6.2.1	2.3.3
444 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.1
445 - Pyruvate metabolism	1.1.1	1.2.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.3
372 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.1
373 - Pyruvate metabolism	3.1.2	1.1.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.3
354 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.1
318 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.1
355 - Pyruvate metabolism	4.4.1	1.2.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.3
319 - Pyruvate metabolism	4.2.3	1.2.1	-.-.-	1.2.2	6.2.1	6.2.1	2.3.3
217 - Phenylalanine metabolism	1.4.3	1.2.1	-.-.-	-.-.-	-.-.-	6.2.1	2.3.1
160 - Glycine, serine and threonine metabolism	1.1.99	1.2.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-
429 - Pyruvate metabolism	-.-.-	1.2.1	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1
205 - Lysine degradation	-.-.-	1.2.4	-.-.-	-.-.-	-.-.-	-.-.-	2.3.1
018 - Glycolysis / Gluconeogenesis	1.8.1	1.2.4	-.-.-	1.2.4	-.-.-	-.-.-	-.-.-
024 - Citrate cycle (TCA cycle)	1.8.1	1.2.4	-.-.-	1.2.4	-.-.-	-.-.-	-.-.-
450 - Pyruvate metabolism	1.8.1	1.2.4	-.-.-	1.2.4	-.-.-	-.-.-	-.-.-
430 - Pyruvate metabolism	-.-.-	1.2.1	-.-.-	-.-.-	-.-.-	-.-.-	2.3.3

***Cluster* 32**

de secuencias: 9 Nombre: >Cluster33 Fitness: 0.405071

	0	1	2	3	4	5	6	7	8
070 - Fatty acid biosynthesis	2.3.1	2.3.1	1.1.1	4.2.1	1.3.1	2.3.1	1.1.1	4.2.1	1.3.1
046 - Pentose and glucuronate interconversions	5.3.1	-.-.-	1.1.1	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	-.-.-
073 - Fatty acid biosynthesis	-.-.-	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.3.1	-.-.-	-.-.-
044 - Pentose and glucuronate interconversions	-.-.-	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	-.-.-
196 - Valine, leucine and isoleucine biosynthesis	4.2.1	-.-.-	-.-.-	4.2.1	1.1.1	2.2.1	1.1.1	-.-.-	-.-.-
054 - Fructose and mannose metabolism	2.7.7	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	-.-.-
213 - Histidine metabolism	-.-.-	-.-.-	-.-.-	4.2.1	1.1.1	-.-.-	1.1.1	-.-.-	6.1.1
212 - Histidine metabolism	-.-.-	-.-.-	-.-.-	4.2.1	-.-.-	-.-.-	1.1.1	-.-.-	6.1.1
021 - Citrate cycle (TCA cycle)	1.1.1	2.3.3	-.-.-	4.2.1	-.-.-	4.2.1	1.1.1	-.-.-	1.1.1

***Cluster* 33**

de secuencias: 16 Nombre: >Cluster34 Fitness: 0.158492

	0	1	2	3	4	5	6	7	8	9	10
425 - Pyruvate metabolism	1.2.1	2.3.3	1.1.99	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
419 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	6.4.1
423 - Pyruvate metabolism	1.2.1	2.3.3	1.1.99	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
417 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.1
424 - Pyruvate metabolism	1.2.1	2.3.3	1.1.99	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
422 - Pyruvate metabolism	1.2.1	2.3.3	1.1.99	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	6.4.1
418 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	4.1.1	2.7.1	1.2.2	2.7.2	3.6.1	6.2.1	6.2.1	2.3.3
416 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	6.4.1
420 - Pyruvate metabolism	1.2.1	2.3.3	1.1.99	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.1
414 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.1
421 - Pyruvate metabolism	1.2.1	2.3.3	1.1.99	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.3
415 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	4.1.1	2.7.1	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.3
410 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	-.-.-	-.-.-	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	6.4.1
408 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	-.-.-	-.-.-	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.1
409 - Pyruvate metabolism	1.2.1	2.3.3	1.1.1	-.-.-	-.-.-	1.2.2	-.-.-	-.-.-	6.2.1	6.2.1	2.3.3
197 - Valine, leucine and isoleucine biosynthesis	1.2.4	2.2.1	1.1.1	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-	-.-.-

***Cluster* 34**

Bibliografía

1. Enzyme Nomenclature. http://www.chem.qmul.ac.uk/iubmb/enzyme/.

2. Metabolic pathway databases. http://www.pathguide.org/.

3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology 215*, 3 (1990), 403-10.

4. Altschul, S.F. Amino acid substitution matrices from an information theoretic perspective. *Journal of molecular biology 219*, 3 (1991), 555-65.

5. Andreas D. Baxevanis, B.F. *Bioinformatics: a practical guide to the analysis of genes and proteins*. John Wiley & Sons, 2001.

6. Arenas-Díaz, E.D. Alineamiento de Múltiples Secuencias Genéticas usando Cómputo Evolutivo. IIMAS, UNAM, Tesis Maestría, 2009.

7. Ay, F., Kahveci, T., and de Crécy-Lagard, V. Consistent alignment of metabolic pathways without abstraction. *Computational systems bioinformatics / Life Sciences Society. Computational Systems Bioinformatics Conference 7*, (2008), 237-48.

8. Basalo, Y.N. Alineamiento Múltiple de Secuencias con T-Coffee : Una Aproximación Paralela. *UAB*, 2009.

9. Berg, J. and Lässig, M. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America 101*, 41 (2004), 14689-94.

10. Cai, L., Juedes, D., and Liakhovitch, E. Evolutionary computation techniques for multiple sequence alignment. *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512)*, , 829-835.

11. Carrillo, H. Lipman, D. The Multiple Sequence Alignment Problem in Biology. *Journal on Applied Mathematics 48*, 5 (2010), 1073-1082.

12. Chapman and Hall. *The Practical Handbook of Genetic Algorithms: Applications.* CRC, 2000.

13. Chen, L., Wang, R., and Zhang, X. *Biomolecular networks: methods and applications in systems biology*. Wiley, 2009.

14. Chen, M. and Hofestaedt, R. An algorithm for linear metabolic pathway alignment. *In silico biology 5*, 2 (2005).

15. Clemente, J.C., Satou, K., and Valiente, G. Finding conserved and non-conserved reactions using a metabolic pathway alignment algorithm. *Genome informatics. International Conference on Genome Informatics 17*, 2 (2006), 46-56.

16. Dandekar, T. and Schuster, S. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *124*, (1999), 115-124.

17. Durbin, R., Eddy, S., and Krogh A. And Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

18. Díaz-Mejia Javier. Una perspectiva de redes sobre la evolución del metabolismo por duplicación génica. IBT, *UNAM, Tesis Doctorado*, 2007.

19. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research 32*, 5 (2004), 1792-7.

20. Fayech, S., Essoussi, N., and Limam, M. Partitioning clustering algorithms for protein sequence data sets. *BioData mining 2*, 1 (2009), 3.

21. Forst, C.V. and Schulten, K. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. *Journal of computational biology : a journal of computational molecular cell biology 6*, 3-4 (1999), 343-60.

22. Gary B. Fogel, D.W. *Evolutionary Computation in Bioinformatics*. 2002.

23. Gerrard, J.A., D, A., and Wells, J.A. Metabolic databases – what next ? *Trends in biochemical sciences 26*, 2 (2001), 137-140.

24. Hariharaputran, S., Töpel, T., Oberwahrenbrock, T., and Hofestädt, R. Alignment of Linear Biochemical Pathways Using Protein Structural Classification. *Nature Precedings*, (2008), 1-5.

25. Henikoff, S. and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America 89*, 22 (1992), 10915-9.

26. Horng, J., Lin, B., and Yang and C-Y, K. A genetic algorithm for multiple sequence alignment. *Proceedings of the GCB*, (2001).

27. John H. Holland. *Adaptation in Natural and Artificial Systems*. Ann Arbor, 1975.

28. Kanehisa, M. and Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research 28*, 1 (2000), 27-30.

29. Kumar, P. Comparative Analysis of k-mean Based Algorithms. *Journal of Computer Science 10*, 4 (2010), 314-318.

30. Lazcano, A. and Miller, S.L. On the origin of metabolic pathways. *Journal of molecular evolution 49*, 4 (1999), 424-31.

31. Lee, Z., Su, S., Chuang, C., and Liu, K. Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. *Applied Soft Computing 8*, 1 (2008), 55-78.

32. Lehninger AL, Nelson DL, C.M. *Principios de Bioquímica*. Barcelona, 1995.

33. Li, Y., de Ridder, D., de Groot, M.J., and Reinders, M.J. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC systems biology 2*, (2008), 111.

34. Liao, L., Kim, S., and Tomb, J. Genome Comparisons Based on Profiles of Metabolic Pathways. *Sixth International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2002)*, (2002), 469-476.

35. Light, S. and Kraulis, P. Network analysis of metabolic enzyme evolution in Escherichia coli. *BMC bioinformatics 5*, (2004), 15.

36. May-Ruíz Gerardo. Alineamiento de Vías Metabólicas. UADY, 2005.

37. Mitchell, M. *Introduction to Genetic Algorithms*. MA, Cambridge, 1996.

38. Needleman, S.B. and Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology 48*, 3 (1970), 443-53.

39. Notredame, C. and Higgins, D.G. Saga: sequence alignment by genetic algorithm. *Nucleic Acids Research 24*, 8 (1996).

40. Notredame, C., Higgins, D.G., and Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology 302*, 1 (2000), 205-17.

41. Notredame, C. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics 3*, 1 (2002), 131-44.

42. Omar, M., Salam, R., Rashid, N., and Abdullah, R. Multiple sequence alignment using genetic algorithm and simulated annealing. *Proceedings. 2004 International Conference on Information and Communication Technologies: From Theory to Applications, 2004.*, , 455-456.

43. Pal, S., Bandyopadhyay, S., and Ray, S. Evolutionary computation in bioinformatics: a review. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 36*, 5 (2006), 601-615.

44. Pearson, W.R. and Lipman, D.J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America 85*, 8 (1988), 2444-8.

45. Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics (Oxford, England) 21*, 16 (2005), 3401-8.

46. Pinter, R.Y. BIOINFORMATICS Alignment of Metabolic Pathways. *Computer*, (2004).

47. Rison, S.C. and Thornton, J.M. Pathway evolution, structurally speaking. *Current opinion in structural biology 12*, 3 (2002), 374-82.

48. Russell, S.J. and Norvig, P. *Artificial Intelligence: A Modern Approach*. 2009.

49. Shannon, C. Prediction and entropy of printed english. *Bell Sys. Tech. 30*, (1951), 50-64.

50. Silva, F.J., Sánchez Pérez, J.M., Gómez Pulido, J.A., and Vega Rodríguez, M.a. AlineaGA—a genetic algorithm with local search optimization for multiple sequence alignment. *Applied Intelligence 32*, 2 (2009), 164-172.

51. Thomas, S., Tapia, L., and Amato, N.M. Protein Folding Core Identification from Rigidity Analysis and Motion Planning. 2008, 7-8.

52. Thompson, J.D., Higgins, D.G., and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific *gap* penalties and weight matrix choice. *Nucleic acids research 22*, 22 (1994), 4673-80.

53. Tohsato, Y. and Nishimura, Y. Metabolic Pathway Alignment Based on Similarity between Chemical Structures. *IPSJ Digital Courier 3*, 1 (2007), 736-745.

54. Tohsato, Y., Matsuda Hideo, and Akihiro, H. A Multiple Alignment Algorithm for Metabolic Pathway Analysis Using Enzyme Hierarchy. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press (2000), 376-383.

55. Waterman, M.S. Identification of Common Molecular Subsequences Identification of Common Molecular Subsequences. (1981), 195-197.

56. Zagordi, O. Statistical Physics Methods in Computational Biology. 2007, 89.

57. Zalzala, A. and Fleming P.J. *Genetic Algorithms in Engineering Systems*. Institute of Electrical Engineers, 1997.