UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Incorporación de Anotaciones Génicas en el Algoritmo de Agrupamiento *MST-kNN*

Daniel Ignacio Pavez Sandoval

TESIS DE GRADO PRESENTADO
EN CONFORMIDAD A LOS REQUISITOS PARA
OBTENER EL GRADO DE MAGÍSTER EN
INGENIERÍA INFORMÁTICA

Profesor Guía: Dr. Mario Inostroza Ponta

SANTIAGO DE CHILE
2013

AGRADECIMIENTOS

Agradezco a todas las personas que estuvieron presentes durante el desarrollo de este trabajo de una u otra forma: a mis familiares (de sangre y políticos), a Camila Azócar (mi amor), a mis amigas y amigos, a mi profesor guía, al proyecto Fondecyt Nº 11121288, y a todos los que depositaron su confianza en mí, me motivaron a seguir adelante de variadas formas, y me dieron la energía para que los años de carrera en la USACH se vieran materializados en este, mi último trabajo universitario.

Al lector, por su interés en este trabajo...

RESUMEN

Dada la necesidad de manejo de enormes volúmenes de datos por parte de los microbiólogos, nace la bioinformática como una forma natural del aporte de herramientas computacionales a la biología. Una técnica utilizada son los algoritmos de agrupamiento que permiten relacionar datos entre sí de acuerdo a la similitud o distancia existente entre ellos. El algoritmo *MST-kNN* es una alternativa basada en grafos de proximidad, que permite relacionar genes entre sí en base a la información que describe su comportamiento bajo las mismas condiciones experimentales.

En la actualidad existen bases de datos de libre acceso que proveen información complementaria de genes, referente a sus características y funcionalidades biológicas asociadas que permiten apoyar el análisis e interpretación de experimentos relacionados al estudio del comportamiento de genes, de la cual no se está utilizando todo su potencial. Si bien los resultados del algoritmo *MST-kNN* ya presentan un grado de coherencia biológica, es deseable que además de utilizar datos del comportamiento de los genes, utilice esa información complementaria con el objetivo de mejorar la coherencia biológica de los grupos que genera.

Este trabajo incorpora la información provista por la base de datos *Gene Ontology* al algoritmo de agrupamiento *MST-kNN* a través del uso de distancias semánticas, estableciendo relaciones entre genes en base a esa información. La solución es probada sobre un conjunto de datos de la levadura *Saccharomyce cerevisiae*, y analizada a través de índices que permiten medir la similitud del comportamiento de los genes de un grupo, y su coherencia biológica. Los resultados son satisfactorios y muestran mejoras de hasta un 93 % en grupos correlacionados según el comportamiento, y además una alta coherencia biológica con mejoras de hasta un 117 %. Al comparar el desarrollo con otra propuesta del estado del arte del problema, se tienen mejoras de hasta un 49 % respecto de la similitud de expresión, y de hasta un 9 % respecto de la coherencia biológica.

Palabras Claves: Bioinformática; Agrupamiento; *MST-kNN*; Anotación Biológica .

ABSTRACT

Given the need of microbiologists to manage large volumes of data, bioinformatics arised as a consequence of the contribution of computational tools for the analysis of this data. A common technique is clustering analysis that aim to generate groups of similar characteristics. The clustering algorithm *MST-kNN* is an alternative based on proximity graphs that relates genes based on the information describing its behavior under the same experimental conditions.

Currently, there are free access databases that provide complementary gene information concerning their associated characteristics and biological functionality. It allows to support the analysis and interpretation of experiments related to genes behavior which is not being used to its full potential. Although the results of the *MST-kNN* algorithm already present a biological coherence degree, it is desirable that in addition to using data from the behavior of genes, it could include biological information in order to improve the biological coherence of the groups that it generates.

This work incorporates the information provided by the GO database to the *MST-kNN* algorithm through the use of semantic distances based on biological information. The solution is tested on a data set of the yeast *Saccharomyce cerevisiae*, and it is analyzed through indexes that measure the similarity of behavior of a group of genes and its biological coherence. The results show that *MST-kNN* algorithm generates groups of genes highly correlated (with improvements of up to 93 %), and also with a high biological coherence (with improvements of up to 117 %). Furthermore, the proposal was compared with another state of the art algorithm. The results are 49 % better in terms of expression similarity and 9 % in terms of biological coherence.

Keywords: Bioinformatics; Clustering; *MST-kNN*; Biological Annotation .

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ALGORITMOS

CAPÍTULO 1. INTRODUCCIÓN

1.1 ANTECEDENTES Y MOTIVACIÓN

Para contextualizar al lector respecto del problema que aborda la presente tesis, se entrega una introducción a los principales conceptos de las áreas científicas involucradas en la investigación, las cuales son: la biología, la informática y la bioinformática. Cuando se trabaja en un área como la bioinformática, no es extrictamente necesario conocer en profundidad todos los conceptos biológicos existentes, pero sí hay que tener claro los involucrados en el trabajo que se lleva a cabo (como por ejemplo, para esta tesis, los conceptos de expresión génica, y anotación o término biológico) además de los conceptos fundamentales y dogma central de la biología. Tal y como lo representa la figura 1.1, la expresión génica se refiere al proceso de producción de ARN por parte de un gen dado, lo que permite estimar el nivel de proteínas que genera, y por tanto, las funciones asociadas a la molécula de ADN de donde se extrajo dicho gen. El ADN es una composición de múltiples copias de una unidad básica llamada nucleótido, el cual se presenta de cuatro formas: *Adenina* (A), *Timina* (T), *Guanina* (G) y *Citosina* (C), formando dos cadenas de pares A-T y G-C enlazados entre sí formando una doble hélice, lo que implica que el ADN puede describirse como una secuencia de nucleótidos y a la vez representarse como una cadena de caracteres “A”, “T”, “G” y “C” de manera que una sub-cadena represente a un gen. Tras un proceso conocido como *transcripción*, la *Timina* es cambiada por el *Uracilo* (U), formando una nueva cadena conocida como ARN, la cual a través de un proceso llamado *traducción* genera una proteína, la cual tendrá una función específica.

Dado lo anterior, un gen no realiza ninguna función por si mismo, sino que mantiene la información necesaria para sintetizar proteínas las cuales realizan las funciones biológicas.

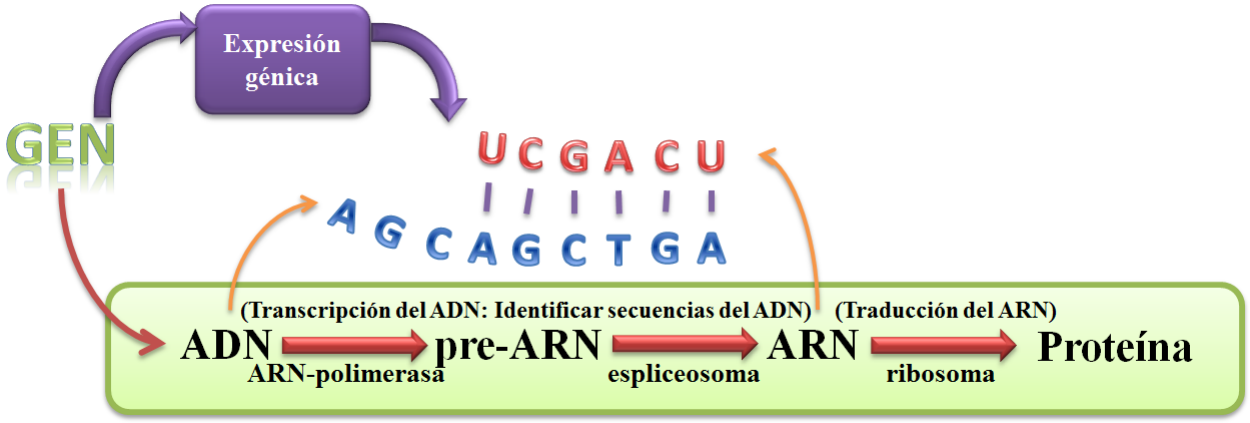

FIGURA 1.1: *Representación del dogma central de la biología, y su relación con el concepto de expresión génica.*

Se debe tener presente que un gen puede producir varias proteínas, y por tanto, tener varias funciones asociadas (COHEN, 2004). Una término o anotación biológica es el elemento en que está hoy en día puesta la atención de la comunidad dedicada a estudios de expresión génica, al ser la información que permite agregar capas de análisis e interpretación a dichos estudios, para extraer un significado biológico (STEIN, 2001). Respecto de una anotación biológica, esta puede estar en tres categorías: a nivel de nucleótido (formando un puente entre la secuencia y la literatura genómica existente), a nivel de proteína (permitiendo la generación de un catálogo de ellas y su correspondiente asignación de funciones putativas) y a nivel de proceso (estableciendo el funcionamiento básico de genes y proteínas, además de sus relaciones).

En los últimos años la biología ha experimentado una evolución veloz en el área de la investigación genómica lo que ha tenido como consecuencia, por ejemplo, los avances respectivos en la medicina. Los biólogos, con ayuda de nuevas tecnologías de alto rendimiento como los de *microarray*, permiten medir el grado de expresión del ARN para una célula en estudio, generando enormes volúmenes de datos con ruido y repeticiones los cuales para ser interpretados requieren del apoyo de herramientas informáticas. Esa mezcla natural entre ambas áreas de la ciencia, da origen a lo que se conoce hoy en día como bioinformática (COHEN, 2004).

Un *microarray* es un *chip* con forma de matriz sobre el cual se depositan varios genes

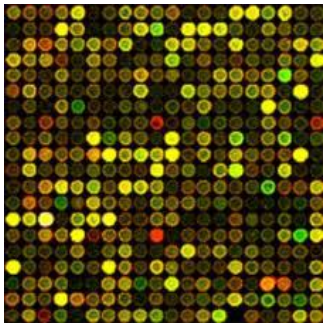

FIGURA 1.2: *Representación del resultado de un experimento de microarray. Las diferentes tonalidades de color e intensidades reflejan la expresión de los genes puestos en el microarray.*

mezclados con una sustancia fluorescente que presenta distintas tonalidades cuando un gen se expresa. Cuando el *chip* lleno de muestras a analizar se somete a condiciones específicas, los genes se expresan o no se expresan dependiendo de su relación con esa condición. La figura 1.2 representa el resultado de un experimento de *microarray*, donde las tonalidades de colores indican la forma y el grado de expresión de un gen frente a condiciones experimentales específicas, permitiendo así su posterior identificación y medición.

Los estudios de expresión génica, con el apoyo de las anotaciones biológicas tienen por delante el desafío de comprender el genoma humano al asociarlo a procesos biológicos (STEIN, 2001), dada su capacidad de estimar las funciones asociadas a una molécula de ADN, por lo que se requiere de un estudio profundo a esa información para interpretar el genoma humano. Una frontera desafiante consiste en aprovechar repositorios de información para el desarrollo de herramientas informáticas y de minería de datos pues, el análisis es un desafío mucho mayor que la generación de la información (GESCHWIND & KONOPKA, 2009). De hecho, presentar el genoma completo (identificación de todos los genes de una cadena de ADN) de alguna especie animal o vegetal es un hito importante pero ambiguo pues, por ejemplo, a finales del año 2000 se dio por terminado el estudio del genoma de la planta *Arabidopsis* cubriendo 115.409.949 pares base con un estimado de 10Mb de regiones no secuenciadas, sin embargo, el largo total del genoma siguió creciendo a razón de 1Mb por año. La importancia de las anotaciones biológicas se hace evidente al establecer una definición más exigente de la completitud de un genoma,

donde se identifica la necesidad de tener extensas anotaciones biológicas de las secuencias en estudio, para extraer así un significado biológico de la investigación (SCHLUETER et al., 2005). Básicamente, se ha de hacer incapié en que no sólo es importante el desarrollo de nuevos métodos y herramientas que permitan generar datos, sino que además aportar en el análisis de ellos para obtener conclusiones que respalden la experimentación, proceso en que es de gran utilidad la información que aportan las anotaciones biológicas.

El último elemento de cual se debe tener claridad, es el aporte de la informática a través de la implementación de algoritmos de agrupamiento basados en métricas numéricas, que pueden resumir y organizar los datos de, por ejemplo, expresión génica permitiendo así identificar características del comportamiento (expresión) de un gen, pues genes con expresión similar probablemente estén relacionados funcionalmente, lo cual se ha podido validar con el conjunto de datos de la levadura *Saccharomyce cerevisiae* (LOCKHART & WINZELER, 2000). Esa estimación de funcionalidades compartidas entre los genes puede ser verificada, de aprovecharse la información de anotaciones biológicas existentes, pues dicha información describe las funciones asociadas a un gen, y por tanto, de ser agrupados los genes con respecto a la similitud que tenga su comportamiento, y además los conjuntos de anotaciones biológicas que los describan, se establece una base fundamentada de las funciones que tienen y que además pueden compartir los genes de un grupo específico.

El algoritmo de agrupamiento *MST-kNN* permite encontrar relaciones existentes en datos de expresión génica, con un grado determinado de coherencia biológica (INOSTROZA-PONTA et al., 2007). Una mejora al mismo se basa en integrar anotaciones biológicas como parámetro de entrada, para que así la evaluación de las relaciones entre genes incorpore una variable de análisis, que permita entregar resultados basados en la información de expresión y en el conocimiento biológico ya adquirido, permitiendo además el análisis del efecto en la calidad de los grupos resultantes generados por el algoritmo, tanto en su similitud de expresión, como en su coherencia biológica. La figura 1.3 representa una posible clasificación de tres algoritmos

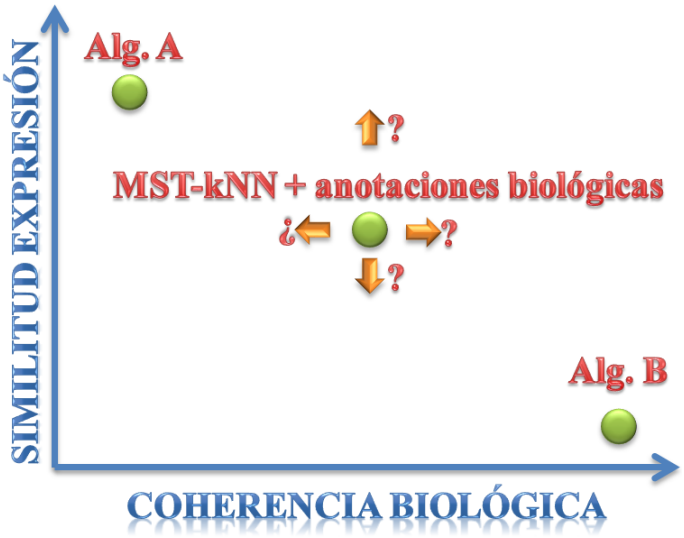

FIGURA 1.3: *Representación del grado de similitud de expresión y coherencia biológica en distintos algoritmos de agrupamiento. Se representan tres algorimos distintos con distintos niveles o grados de similitud de expresión génica y coherencia biológica.*

de agrupamiento para datos de expresión génica, donde se aprecia que todos poseen un nivel determinado de similitud de expresión (eje de las ordenadas), y de coherencia biológica (eje de las abscisas) de manera que "*Alg. A*" posee un nivel de similitud de expresión mayor al de "*Alg. B*" (con un grado de coherencia biológica menor), mientras que "*Alg. B*" posee un grado de coherencia biológica mayor que el de "*Alg. A*" (con un grado de similitud de expresión menor). Por tanto, al incorporar las anotaciones biológicas al algoritmo *MST-kNN* se podrá analizar el efecto de la calidad de los grupos medida en relación tanto de su nivel de similitud de expresión, como de su coherencia biológica.

Actualmente hay algunos desarrollos relacionados a la incorporación de anotaciones biológicas a algoritmos de agrupamiento, para con ello obtener grupos de genes con un mayor grado de coherencia biológica. Se tiene, por ejemplo, una incorporación de información de *Gene Ontology* (GO, 2013), a los algoritmos de agrupamiento *Super Paramagnetic Clustering* (CHERNOMORETZ, 2010), *SemiSuperviced Possibilistic Clustering Algorithm* (MARAZIOTIS et al., 2012), y a un algoritmo de agrupamiento no supervisado (VERBANCK et al., 2013). Los desarrollos mencionados se describen con mayor detalle en la sección 2.4.

1.2 DESCRIPCIÓN DEL PROBLEMA

En el presente trabajo se plantea, dada la capacidad de agrupamiento que posee el algoritmo *MST-kNN*, añadir como parámetro del mismo anotaciones biológicas de un organismo específico. El problema se basa, entonces, en encontrar una manera de incorporar al proceso de agrupamiento basado en expresión génica, la información de anotaciones biológicas existentes en bases de datos de libre disposición, para así generar grupos de genes que estén biológicamente relacionados y que además posean patrones de expresión génica similares, es decir, que las descripciones funcionales de los genes pertenecientes a un grupo sean coherentes, y que además su comportamiento bajo las mismas condiciones experimentales sea similar.

1.3 SOLUCIÓN PROPUESTA

La solución propuesta es un modelo matemático de transformación del conocimiento biológico estático a una matriz de distancias en base a similitudes semánticas, que permita combinar su contenido al de una matriz generada a partir de datos de expresión génica, de manera que una misma estructura contemple ambos tipos de datos y pueda ser incorporada al algoritmo de agrupamiento *MST-kNN*, y en rigor, a cualquier algoritmo de agrupamiento basado en grafos.

La solución está entonces compuesta por un modelo procedural respaldado por fundamentos teóricos, que describen la transformación de información semántica estática a una matriz de distancias, que es luego combinada con una matriz generada a partir de datos numéricos para ser incorporada como parámetro de entrada al algoritmo de agrupamiento *MST-kNN*, y analizar los resultados a través de índices de validación coherentes con la incorporación de información. La solución también debe considerar la implementación computacional del modelo procedural propuesto, para validar de forma empírica los resultados obtenidos que

corresponden a un grafo compuesto de un número determinado de partes conexas o grupos, que consideran expresión que describe el comportamiento de un gen, y además la información estática que caracteriza biológica y funcionalmente al mismo gen, y que es un complemento de análisis importante al asignar un sentido biológico a datos que de otra manera no representan una mayor cantidad de información por sí mismos (MEYER et al., 2003). El algoritmo *MST-kNN* no presenta cambios en su salida, por lo que no se limita su integración a productos de *software* como *QAPGrid*, que permite una representación visual de los resultados (INOSTROZA-PONTA et al., 2011). La eficiencia actual del algoritmo *MST-kNN* no se verá alterada, manteniendo un orden de complejidad de $\theta(n) = n^2 \log(n)$, donde la variable n representa al total de elementos a agrupar, por ejemplo, al total de genes en estudio (PALMA, 2011).

La solución pretende aumentar el grado de información incluida en las asociaciones biológicas entregadas por el algoritmo *MST-kNN*, al considerar otras variables como la ubicación cromosómica y subcelular de la expresión génica, características del gen y las funciones biológicas asociadas a él, lo que facilita el manejo de los grandes volúmenes de datos, al identificar grupos de menor volumen que sean buenos candidatos para un posterior análisis, implicando así la generación de nuevo conocimiento biológico.

1.4 OBJETIVOS Y ALCANCE DEL PROYECTO

1.4.1 Objetivo general

Dado un conjunto de genes, el objetivo es obtener un modelo de datos que incorpore anotaciones biológicas a datos de expresión génica, en una fase previa al agrupamiento generado por *MST-kNN*, para que, a partir de los datos de expresión génica y de anotaciones biológicas de dichos genes, sea capaz de asignarlos a grupos biológicamente relacionados y con patrones

de expresión génica similares, modificando así el grado de información y coherencia biológica en experimentos cuyo objetivo sea establecer similitudes entre genes.

1.4.2 Objetivos específicos

Para la consecución del objetivo general, se plantean las siguientes metas intermedias:

1. Conocer el funcionamiento del algoritmo *MST-kNN* a nivel teórico y de implementación, para identificar la estructura de datos que compone su funcionamiento.

2. Identificar las bases de datos públicas de anotaciones biológicas, precisar sus características y revisar en la literatura métodos para representar el contenido extraído de ellas.

3. Desarrollar un modelo conceptual que incorpore las anotaciones biológicas extraídas a datos de expresión génica, como una fase previa al algoritmo de agrupamiento *MST-kNN*.

4. Validar empíricamente el modelo propuesto, sobre el conjunto de datos de la levadura *Saccharomyces cerevisiae* (EISEN et al., 1998) y las fuentes de anotaciones génicas almacenadas en *Gene Ontology* (GO, 2013).

1.4.3 Alcances

Una característica del algoritmo de agrupamiento *MST-kNN*, es que puede ser aplicado en diferentes disciplinas, siempre que se puedan modelar los datos de tal forma que sean compatibles con la entrada requerida. A pesar de esa flexibilidad, para este trabajo sólo se utiliza la relacionada a investigación biológica, específicamente al estudio de expresión génica, para la levadura *Saccharomyce cerevisiae*. Por otro lado, las anotaciones biológicas de la levadura son extraídas sólo de la base de datos de ontologías biológicas GO. El modelo de incorporación del

conocimiento biológico a los datos de expresión será aplicado sólo al algoritmo de agrupamiento *MST-kNN*, sin considerar su utilización en otro algoritmo de agrupamiento basado en grafos. Por último, queda fuera de este trabajo la integración del modelo, a un algoritmo de visualización como *QAPGrid* y también la interpretación biológica de los resultados.

1.5 METODOLOGÍA Y HERRAMIENTAS UTILIZADAS

1.5.1 Metodología

La presente investigación se divide en dos etapas, la primera es el proceso investigativo de carácter científico asociado a la solución de un problema del área de la bioinformática, y la segunda, es el desarrollo de una herramienta computacional que implemente a nivel de *Software* la solución encontrada al problema, para probar y validar el desarrollo teórico. Es por ello que son consideradas dos metodologías, una directamente relacionada al proceso investigativo y la otra a la generación del *Software* (cuyas descripciones complementarias se presentan en el apéndice A.1).

Referente al trabajo investigativo, la metodología seleccionada es el método científico, dada su naturaleza de apoyo a la integración a un área en la cual no se tiene conocimiento o experiencia. La descripción de las etapas, y su relación con los objetivos del proyecto se presenta a continuación:

- **Observación**: Revisión bibliográfica y estudio del marco teórico relevante, que permite justificar en términos matemáticos y científicos las decisiones respecto de la solución propuesta, formas de validación, entre otras problemáticas propias de la investigación. Esta primera fase se ve materializada al estudiar el área de la bioinformática, la teoría de grafos, el algoritmo *MST-kNN*, medidas de similitud semántica y otros campos necesarios para el desarrollo de la tesis.

- **Inducción**: Se establece una relación entre la teoría estudiada en la etapa anterior, y los lineamientos que se han de seguir para establecer de forma clara tanto la pregunta de investigación, como la hipótesis y propuesta de solución.

- **Hipótesis**: Corresponde a la formulación de una suposición o idea, que permite describir deducciones a partir de ella.

- **Prueba o refutación de la hipótesis**: Sometimiento a pruebas empíricas de la hipótesis, con el objetivo de determinar si es válida o no. El desarrollo de esta etapa va acompañada con un diseño experimental que permite responder a la pregunta de investigación, o problemática que se pretende solucionar.

- **Conclusiones**: Etapa final de la metodología, luego de realizadas las pruebas y contrastados los resultados con los esperados por la hipótesis. Permite realizar un análisis y autocrítica de la validez de los resultados, verificándolos e interpretándolos.

Dado que la metodología implica el planteamiento de una hipótesis acorde a lo descrito en la sección 1.2, a continuación se presenta la propuesta que se desarrolla en la tesis:

"Dada la capacidad de agrupamiento del algoritmo *MST-kNN*, es posible añadir como parámetro del mismo, anotaciones biológicas de la levadura *Saccharomyces cerevisiae* disponibles en bases de datos de libre acceso, para encontrar grupos que estén asociados tanto a nivel de la expresión génica del organismo, como de sus anotaciones biológicas".

La etapa de pruebas que permitirá llegar al objetivo presentado en la sección 1.4.1 consta en la comparación de la solución obtenida con el algoritmo original que no incluye anotaciones biológicas. Lo anterior tiene sentido, debido a que, como se observa en la figura 1.3, los datos de expresión por sí mismos también incorporan un grado de información biológica. Se debe realizar una validación y prueba de la solución que genere un contraste que tenga sentido y además permita validar las capacidades de la solución propuesta, es por ello que se compara además con otro desarrollo encontrado en la literatura relacionado al estado del arte del problema atacado.

Referente a la metodología asociada al desarrollo del *Software*, dadas las características de la aplicación a desarrollar, la que mejor se adapta a las necesidades de la tesis es *R.A.D.* (*Rapid Application Development* o Desarrollo Rápido para Aplicaciones), la cual en base a iteraciones genera prototipos sin dar mayor énfasis en la elaboración de documentación asociada al desarrollo. La metodología es adecuada, dado que permitirá la realización de pruebas parciales al sistema, validando de inmediato el avance que se está generando, y porque para el desarrollo de la presente tesis el *Software* es sólo una herramienta que permitirá validar y respaldar el desarrollo teórico, sin ser un producto del trabajo mismo.

1.5.2 Herramientas de desarrollo

El lenguaje de programación en que se desarrollan las implementaciones es JAVA, utilizando el *kit* de desarrollo (JDK) en su versión 7, a través del entorno de desarrollo integrado libre **NetBeans**, en su versión 7.3.1. Las herramientas utilizadas en el desarrollo de documentación y construcción de la tesis, son las provistas por el sistema de composición de textos de *Software* libre LaTeX, a través del Servicio *Web ShareLaTeX* que permite la edición, mantención y compartición de documentos en forma colaborativa a través de la *Web*. Las herramientas provistas por *Microsoft Office* (*Word*, *PowerPoint*, *Excel* y *Project*) son utilizadas para permitir la lectura de otros documentos que contribuyan al desarrollo de la tesis, para construir algún tipo de diagrama o gráfico, facilitar el cálculo de datos tabulados o bien para mantener información temporal durante el desarrollo. La máquina física donde se llevó a cabo la implementación del *Software* y se realizaron las pruebas, es un *Laptop* con Sistema Operativo *Windows 7*, y *Ubuntu* en su versión 12.04, ambos con arquitectura de 64 bits. El procesador corresponde a un *Intel Core i3*, modelo M350 con frecuencia de 2.27Ghz, y además cuenta con 4GB de memoria RAM, y un disco duro de 320GB de capacidad.

El desarrollo de la etapa investigativa se llevó a cabo en el Departamento de Ingeniería

Informática (DIINF) de la Universidad de Santiago de Chile (USACH), en la sala de Informática Colaborativa, y además en el lugar de residencia del estudiante. En principio, se llevaron a cabo video conferencias con el profesor guía, Dr. Mario Inostroza-Ponta, a través de la plataforma *Web Google Plus Hangouts*, para posteriormente realizar reuniones de coordinación en el DIINF de la USACH.

1.6 ORGANIZACIÓN DEL DOCUMENTO

Respecto del orden y contenidos de la tesis, el capítulo número dos entrega la información relacionada a los materiales, herramientas, y conceptos teóricos de interés para el desarrollo. Luego que se tienen los elementos teóricos fundamentales, el capítulo tres da a conocer cómo éstos fueron utilizados, modelados y modificados para dar solución al problema fundamental expuesto en la sección 1.2.

Una vez descrita la solución desarrollada, el capítulo número cuatro presenta los resultados que se obtienen luego de la experimentación (la definición de los experimentos realizados se lleva a cabo en el mismo capítulo), para con ello establecer las principales conclusiones que se presentan en el capítulo número cinco. El documento además contiene tres apéndices de información complementaria asociada a las metodologías y algoritmos utilizados, a los elementos propios del área de la bioinformática, y también a información de apoyo a los resultados obtenidos.

CAPÍTULO 2. MARCO TEÓRICO

El presente capítulo tiene por objetivo dar a conocer los contenidos que son el sustento teórico de la solución propuesta e implementada. Los temas que acá se encuentran son fundamentales para comprender en su totalidad tanto la solución, como las pruebas y análisis, pues en ellos se basan decisiones que influyen directamente en el desarrollo.

2.1 EXPRESIÓN GÉNICA

Los datos de expresión de un gen permiten describir su comportamiento bajo condiciones experimentales conocidas, representando así su transformación a una o más proteínas las cuales cumplirán una función específica. Se puede, entonces, tener varias muestras (de experimentos diferentes) relacionadas a los valores de expresión para un conjunto de genes, lo que permite comparar sus comportamientos al medir en qué grado se expresan (o correlacionan) frente a una misma condición experimental. Considerar por ejemplo tres genes *Gen1*, *Gen2* y *Gen3* (tablar 2.1), los cuales son sometidos a diez experimentos diferentes. Así pues, el conjunto de valores $\{-1, -0.3, -0.12, -0.08, 0, 0.7, 0.1, 0.4, 0.7, 1\}$ corresponde al perfil de expresión génica del *Gen1*, para el conjunto de muestras $\{M0, M1, M2, M3, M4, M5, M6, M7, M8, M9\}$. Se tiene entonces treinta valores de expresión los cuales se grafican en la indica la figura 2.1, lo que permite estudiar qué tan similar es el comportamiento entre dos genes bajo las mismas condiciones tras calcular la correlación entre sus comportamientos, tal y como se describe en la sección 2.1.1.

TABLA 2.1: Ejemplo de diez valores de expresión para tres genes.

Gen	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9
Gen1	-1	-0.3	-0.12	-0.08	0	0.7	0.1	0.4	0.7	1
Gen2	1	0.3	0.12	0.08	0	-0.7	-0.1	-0.4	-0.7	-1
Gen3	0	0.7	0.88	0.92	1	1.7	1.1	1.4	1.7	2

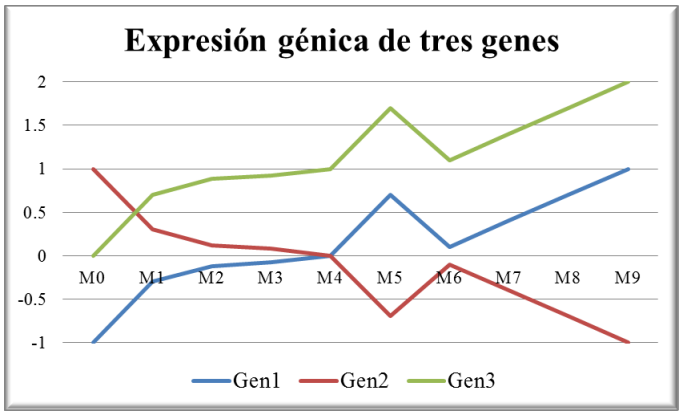

FIGURA 2.1: *Ejemplo del comportamiento de tres genes, visto a través de sus valores de expresión de diez muestras diferentes, los cuales corresponden a los de la tabla 2.1.*

2.1.1 Distancia entre perfiles de expresión

En la selección de una métrica de distancia entre perfiles de expresión génica, una basada en el coeficiente de correlación de *Pearson* (ρ, el cual se observa en la ecuación 2.1) da mejores resultados que una basada en distancia *Euclidea* cuando lo que se desea medir es el correlación de los genes (INOSTROZA-PONTA, 2008). El coeficiente ρ, al ser independiente de la magnitud permite evaluar qué tan similar se comportan los genes de acuerdo a su perfil de expresión, aunque tiene como desventaja no ser robusta para valores atípicos, lo que significa que si dos genes coinciden en una cima o valle del comportamiento de una muestra, la correlación será dominada por esa característica, aunque el resto del comportamiento sea diferente. El coeficiente ρ entrega valores dentro del intervalo $[-1, 1]$, tomando el valor -1 cuando los elementos en comparación están totalmente anti-correlacionados (ejemplo, los genes *Gen1* o *Gen3*, con respecto al *Gen2* según la figura 2.1), el valor cero cuando no hay correlación y el

valor uno cuando están completamente correlacionados (ejemplo, el *Gen1* con respecto al *Gen3* según la figura 2.1).

Al calcular $\left(1-\rho_{g_x g_y}\right)$ se tiene el análogo de la similitud en forma de distancia, ya que cuando hay correlación la distancia tiende a cero, y si hay anti-correlación tiende a dos (distancia máxima). Una variante es considerar a la sobre-expresión y sub-expresión (correlación y anticorrelación respectivame) entre genes como un comportamiento correlacionado, es decir, se co-expresan o tienen un comportamiento similar (aunque sea de diferente magnitud) bajo las mismas condiciones; esa variante se consigue al calcular $\left(1-\left|\rho_{g_x g_y}\right|\right)$ con intervalo de valores $[0,1]$, donde valores cercanos a uno indican que los genes en evaluación no se expresan de forma similar, y valores cercanos a cero indican que se co-expresan.

$$\rho_{g_x g_y} = \frac{\sum_{i=1}^{m} g_x(i)\, g_y(i) - \frac{\sum_{i=1}^{m} g_x(i) \sum_{i=1}^{m} g_y(i)}{m}}{\sqrt{\left(\sum_{i=1}^{m} g_x(i)^2 - \frac{\left(\sum_{i=1}^{m} g_x(i)\right)^2}{m}\right)\left(\sum_{i=1}^{m} g_y(i)^2 - \frac{\left(\sum_{i=1}^{m} g_y(i)\right)^2}{m}\right)}} \tag{2.1}$$

2.1.2 Índice de coexpresión

Con el objetivo de analizar si un conjunto de genes tiene comportamientos similares bajo las mismas condiciones experimentales, Marie Verbanck propone el índice empírico expuesto en la ecuación 2.2 llamado *Índice de Co-expresión* (VERBANCK et al., 2013), desde ahora *IC*, el cual se aplica a un grupo de genes (C_g) haciendo uso de su cardinalidad ($\#\left(C_g\right)$), y M muestras de valores de expresión (E_{mg} valor de expresión del gen g para la muestra m) a los cuales se les calcula tanto el valor promedio de todas las muestras ($\bar{E}_g$) como su desviación estándar

(σ_{E_g}). IC tiene valores en el intervalo $[-1, 1]$, de manera que -1 indica que los genes del grupo están anti-correlacionados, uno que están correlacionados y cero que no hay correlación entre ellos. Este índice será aplicado a experimentos que utilicen la ecuación 2.1 para establecer la correlación entre genes en base a sus perfiles de expresión.

$$IC\left(C_g\right) = \frac{2}{\#\left(C_g\right)\left(\#\left(C_g\right)-1\right)} \sum_{g|g\in C_g} \sum_{g'|g'\in C_g, g'>g} \frac{1}{M} \sum_{i=1}^{M} \frac{E_{mg} - \bar{E}_g}{\sigma_{E_g}} \frac{E_{mg'} - \bar{E}_{g'}}{\sigma_{E_{g'}}} \tag{2.2}$$

En base al índice IC expuesto en la ecuación 2.2, considerar:

- $f_{Card} = \frac{2}{\#(C_g)(\#(C_g)-1)}$
- $\rho_{g,g'} = \sum_{g'|g'\in C_g, g'>g} \frac{1}{M} \sum_{i=1}^{M} \frac{E_{ig} - \bar{E}_g}{\sigma_{E_g}} \frac{E_{ig'} - \bar{E}_{g'}}{\sigma_{E_{g'}}}$

Con lo anterior, se propone una variación de IC (ecuación 2.3) con intervalo de valores $[0, 1]$, que considera que dos genes están correlacionados tanto si sus perfiles de expresión están sobre-expresados como sub-expresados (valores cercanos a uno) y será aplicada a experimentos que consideren ese mismo comportamiento para los genes en estudio.

$$IC_{Abs}\left(C_g\right) = f_{Card} \sum_{g|g\in C_g} \left(|\rho_{g,g'}|\right) \tag{2.3}$$

2.2 ANOTACIÓN BIOLÓGICA

2.2.1 Proyecto *Gene Ontology* (GO)

Dada la existencia de más de un laboratorio dedicado a estudios de notación genómica, cada uno con su propia nomenclatura, se hace difícil la búsqueda de información específica de

alguna especie en particular. En el año 1998, y como una forma de hacer frente al mencionado problema, comienza el proyecto GO, el cual mantiene en una única base de datos la información de diversas fuentes como *FlyBase*, *Saccharomyces Genome Database* y *Mouse Genome Database* (de los organismos mosca *Drosophila*, levadura y ratón, respectivamente). El proyecto, de carácter colaborativo sin fines de lucro, mantiene descripciones de productos génicos de variadas especies de plantas y animales. Actualmente es un consorcio conformado por veinte miembros y seis colaboradores (listados en el apéndice B.1.1) que no pretenden imponer un estándar, sino que cooperar para llegar a un consenso siendo un aporte a los procesos investigativos, dada la facilidad del conocimiento de quedar rezagado.

GO mantiene tres ontologías estructuradas que permiten, por ejemplo, describir funcionalmente a un gen. Las ontologías se mantienen de forma independiente, lo que facilita la búsqueda de información con diferentes niveles de especificidad. En particular, las tres ontologías de GO son:

- **Componente celular**: Mantiene información de las partes que conforman una célula, o a su entorno extracelular.
- **Proceso biológico**: Conjunto de operaciones con inicio y término definidos que describen el funcionamiento de células, tejidos, órganos u organismos.
- **Función molecular**: Actividades que lleva a cabo el producto de un gen a nivel molecular.

La estructura de las tres ontologías descritas es la de un grafo acíclico dirigido (*directed acyclic graph*, DAG), lo que implica que (visto de manera jerárquica) existe un nodo raíz, y cada descendiente puede tener uno o más ancestros. Un nodo de una ontología (que representa a un término biológico) puede además relacionarse con uno o más nodos de otra ontología. En la figura 2.2 se ve un ejemplo de seis anotaciones biológicas pertenecientes a la ontología "proceso biológico" (*biological process*), los cuales son presentadas a través de la herramienta *Web* oficial

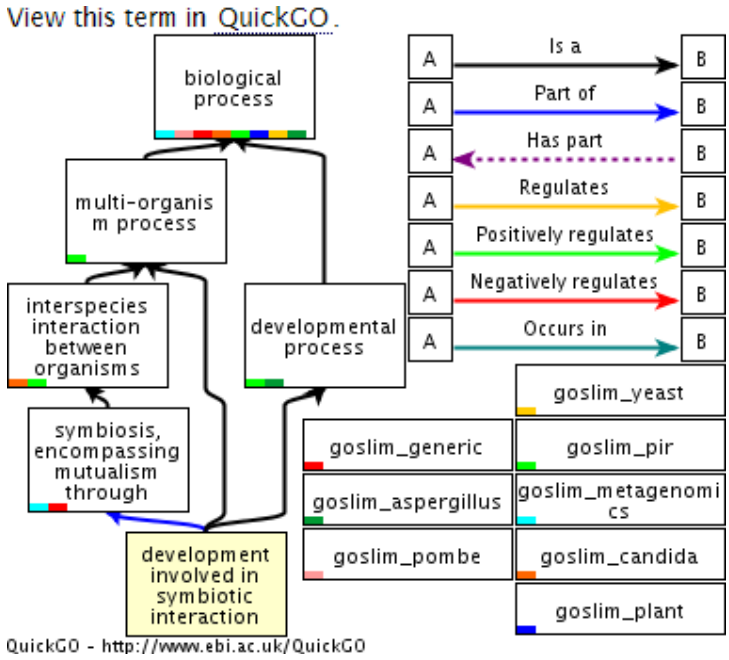

FIGURA 2.2: *Extracto de la ontología "proceso biológico", donde se representa la estructura de grafo acíclico dirigido (DAG).*

del proyecto para extraer información de GO, *AmiGO*. Además de servicios *Web*, hay *Softwares* instalables como por ejemplo la herramienta *OBO-Edit*. Cabe mencionar que GO no mantiene información de perfiles de expresión génica, sino que describe su comportamiento en un contexto celular, también omite procesos, funciones y componentes de mutaciones o enfermedades por no representar la función normal de un gen. Además del acceso vía herramientas, GO permite la descarga de la base de datos a través de las siguientes opciones:

1. ***termdb***: De actualización diaria, corresponde a los datos relacionados a las ontologías, definiciones y mapeos a otras bases de datos.

2. ***assocdb***: De actualización semanal, contiene a *termdb*, y todas las anotaciones de productos de genes manuales y anotaciones inferidas electrónicamente de todas las bases de datos menos de *UniProt Knowledgebase* (UniProtKB, 2013), debido al tamaño de los conjuntos de datos.

3. ***seqdb***: De actualización semanal, contiene a *assocdb*, más las secuencias de proteínas para la mayoría de los productos génicos.

4. ***full GO database***: De actualización mensual, contiene a *seqdb* más un manual y anotaciones inferidas electrónicamente.

Hay que tener presente que cada término biológico de GO es el resultado de numerosos experimentos que certifican que dos términos están relacionados entre sí de forma específica. Dado lo anterior, la información contenida en GO contiene implícitamente un respaldo teórico valioso, que actualmente no es aprovechado en su totalidad por la comunidad científica encargada de estudiar las similitudes entre genes en base a los perfiles de expresión génica. Cada gen_i asociado a un término es información validada y certera referente a qué función tiene o cumple dicho gen, y más aun, identifica a una función relacionada a otras funciones que pueden potencialmente estar relacionadas al gen_i, o a otros genes relacionados a él. Más detalles respecto de la base de datos de GO, se encuentran en el apéndice B.1.2. En particular, para el presente trabajo se utilizaron las anotaciones biológicas de los genes de la levadura *Saccharomyces cerevisiae* (EISEN et al., 1998), para establecer sus perfiles funcionales, que representan a las anotaciones biológicas que describen funcionalmente a un gen (VERBANCK et al., 2013).

2.2.2 Distancia entre perfiles funcionales

La presente sección contesta a la pregunta, expuesta en la sección 1.2, de ¿cómo incorporar al agrupamiento basado en expresión génica, anotaciones biológicas? El camino seleccionado para contestar la interrogante, es calcular qué tan similares son dos genes, considerando los perfiles funcionales que los describen (conjunto de anotaciones biológicas asociados a ellos) y además sus perfiles de expresión (comportamiento de ambos genes en las mismas condiciones experimentales).

La solución se divide en tres etapas: (1) medir la distancia entre pares de anotaciones biológicas, (2) en base a los conjuntos de anotaciones biológicas de los genes medir la distancia entre pares de ellos (considerando las distancias entre pares de anotaciones biológicas ya calculadas), y (3) combinar una matriz de distancia entre genes en base a perfiles de expresión (lo que ya existe) con una matriz de distancia entre genes en base al conocimiento biológico (lo que se propone). En la presente sección se aborda la primera etapa, que consiste en calcular la distancia entre pares de anotaciones biológicas.

Dada la estructura en que se distribuyen los términos biológicos en GO (graficado en la figura 2.2), tiene sentido calcular las distancias semánticas entre los términos a partir de su ubicación dentro del DAG. Entendiendo como nodo a un término biológico, y como arista a la relación entre dos términos, la medición de la distancia semántica puede hacerse al contar el número de aristas que separan a dos nodos dentro de grafo (enfoque basado en las aristas); calculando el "Contenido de Información" (desde ahora, C.I.) de los nodos, de sus descendientes y sus ancestros (enfoque basado en los nodos), o bien combinando ambos enfoques (SHENOY et al., 2012).

En ambos enfoques se utiliza el concepto de ancestro común mínimo (desde ahora, A.C.M.), ya sea para calcular el camino mínimo existente entre dos nodos (concepto de Largo del Camino Mínimo, L.C.M.), o como nodo de referencia para el cálculo del C.I. (concepto que se explica más adelante). Dado que el enfoque basado en las aristas utiliza el concepto de L.C.M. el cual asume una distribución uniforme de tanto nodos como aristas dentro del grafo, es cuestionada su efectividad dentro de las ontologías biológicas, por la ausencia de esas características (BENABDERRAHMANE et al., 2010).

El A.C.M. de dos nodos N_X y N_Y, representa al nodo más específico (ubicado en el nivel más alejado de la raíz posible) que es ancestro de N_X y N_Y a la vez ($N_{A.C.M.}$), mientras que el L.C.M. entre los nodos N_X y N_Y, es la distancia mínima entre el nodo N_X y $N_{A.C.M.}$ (denotémosla como $d_{N_X - N_{A.C.M.}}$), sumada a la distancia mínima entre N_Y y $N_{A.C.M.}$

($d_{N_Y - N_{A.C.M.}}$), es decir $L.C.M. = d_{N_X - N_{A.C.M.}} + d_{N_Y - N_{A.C.M.}}$. Por otro lado, el concepto de C.I. permite evitar la asumpción de una distribución uniforme de nodos y aristas en la taxonomía, al considerar que dos nodos tienen información en común a través de la probabilidad de encontrar una instancia de ellos dentro de la taxonomía (frecuencia del nodo dividido por el total de elementos de la taxonomía). Con lo anterior, un nodo N_A tiene probabilidad $p(N_A) = \frac{frec(N_A)}{totalElementosTaxon}$, con valores posibles en el intervalo $[0, 1]$. En las ontologías biológicas la probabilidad de una instancia de una anotación biológica está sujeta a su frecuencia de aparición, que considera (además de la aparición misma de la anotación biológica) a todos los descendientes que posee, es decir, la información de una anotación biológica N_A está en él y en todos sus descendientes. La argumentación estándar de la teoría de información, indica que se ha de calcular como el logaritmo negativo de la probabilidad $-\log(p(N_A))$, por lo tanto si N_A es ancestro de N_B, $p(N_A) > p(N_B)$, y por tanto el contenido de información de N_B es mayor que el de N_A. Con lo anterior, se entiende que el contenido de información del nodo raíz es cero, por tener probabilidad igual a uno. En definitiva, a medida que el nodo en evaluación es más específico (su probabilidad disminuye) su contenido de información aumenta, y viceversa.

Para ejemplificar los conceptos de A.C.M., L.C.M. y C.I. expuestos anteriormente, se considera una estructura como la expuesta en la figura 2.3 para indicar los valores de A.C.M., L.C.M., frecuencia, probabilidad y C.I. de algunos nodos presentes en la tabla 2.2.

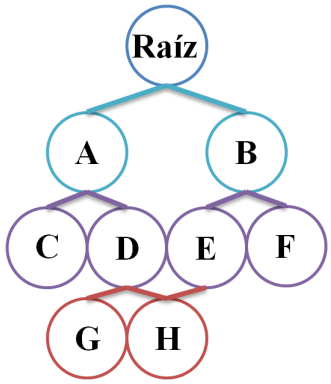

FIGURA 2.3: *Ejemplo de DAG con nueve nodos y nueve relaciones entre ellos.*

TABLA 2.2: Ejemplo del cálculo de algunos valores de A.C.M., L.C.M., y de los valores de Frecuencia, Probabilidad y C.I. de los nodos en base a la estructura presente en 2.3.

Nodos	A.C.M.	L.C.M.	Nodo	Frecuencia	Probabilidad	Contenido Información
Raíz-H	Raíz	3	Raíz	9	1	0
A-D	A	1	A	5	0.5556	0.2553
B-C	Raíz	3	B	4	0.4444	0.3522
C-G	A	3	C	1	0.1111	0.9542
D-C	A	2	D	3	0.3333	0.4771
E-F	B	2	E	2	0.2222	0.6532
F-H	B	3	F	1	0.1111	0.9542
G-Raíz	Raíz	3	G	1	0.1111	0.9542
H-C	A	3	H	1	0.1111	0.9542

La figura 2.4 es una representación de los elementos necesarios para el cálculo de una medida de similitud semántica basada en las aristas, donde el elemento $N3$ representa al A.C.M. entre los nodos $N1$ y $N2$, mientras que $D1$ representa a la distancia entre el elemento $N1$ y $N3$ y $D2$ la distancia entre $N2$ y $N3$. Por último, $D3$ representa la mínima distancia entre $N3$ y el nodo raíz "*Raíz*".

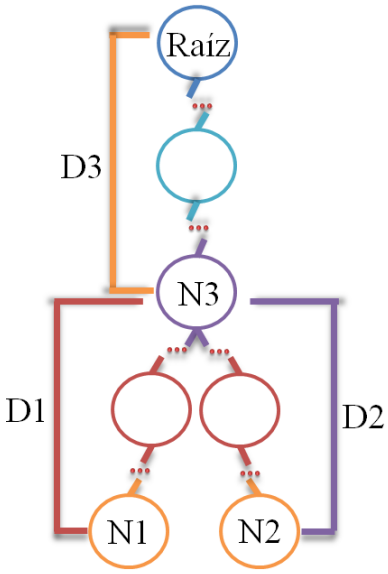

FIGURA 2.4: *Estructura jerárquica propuesta en (WU & PALMER, 1994). $N3$ representa al A.C.M. entre $N1$ y $N2$, $D1$ la distancia entre el elemento $N1$ y $N3$ y $D2$ la distancia entre $N2$ y $N3$. $D3$ representa la mínima distancia entre $N3$ y el nodo raíz "Raíz".*

2.2.2.1 Similitud de Wu-Palmer

Con el objetivo de mejorar una máquina de traducción inglés-chino, en (WU & PALMER, 1994) se propone una representación semántica de verbos a través de una estructura jerárquica que permite calcular la similitud semántica entre pares de ellos, con el fin de mejorar la selección léxicamente correcta en el proceso de traducción, al considerar que la similitud de dos verbos se define por cuán cercanos estén relacionados dentro de la jerarquía. La medida de similitud (que utiliza los elementos descritos en la figura 2.4) se presenta en la ecuación 2.4.

$$Sim_{wp}\left(N_A, N_B\right) = \frac{2 * D3}{D1 + D2 + (2 * D3)} \tag{2.4}$$

2.2.2.2 Similitud de T.B.K. (Slimani, Yaghlaney y Mellouli)

El enfoque propuesto en la sección 2.2.2.1 tiene como desventaja el hecho de considerar que dos nodos de una misma jerarquía tienen valores de similitud menores que los de una misma $vecindad_{var}$ (SHENOY et al., 2012), considerando que dos nodos están en una misma jerarquía cuando están en un mismo nivel de profundidad dentro del árbol como respecto al nodo raíz, y en una misma $vecindad_{var}$ (variación del concepto de vecindad) cuando no están en una misma jerarquía, y además uno no es ancestro o descendiente del otro, por ejemplo, en la figura 2.3, los nodos A y B, o G y H están en una misma jerarquía, mientras que los nodos C y B, A y F, o C y H están en una misma $vecindad_{var}$. Con el objetivo de equilibrar esa consideración, se propone una medida de similitud basada en el conteo de aristas que es una variación de la presentada por (WU & PALMER, 1994), manteniendo así sus ventajas de simplicidad del cálculo y alto rendimiento, al agregar un factor que penaliza la similitud entre dos nodos de una misma $vecindad_{var}$ (SLIMANI et al., 2008). El enfoque se crea para enfrentar el problema de

identificación semántica asegurando que $Sim\,(N_X, N_Y) \leq Sim\,(N_X, N_W)$, cuando N_Y pertenece a la misma $vecindad_{var}$ que N_X, y N_W a la misma jerarquía de N_X. La propuesta de (SLIMANI et al., 2008) se presenta en la ecuación 2.5.

$$Sim_{tbk}\,(N_A, N_B) = Sim_{wp} \times F\,(N_A, N_B) \tag{2.5}$$

Donde:

$$F\,(N_A, N_B) = \left((1-\lambda)\,(\text{mín}(D1 + D3, D2 + D3)) + \lambda\,(|D1 - D2| + 1)^{-1}\right) \tag{2.6}$$

En la ecuación 2.6 el coeficiente λ toma el valor cero si los dos nodos están en la misma jerarquía, y el valor uno si están en la misma $vecindad_{var}$. Si N_A es ancestro de N_B, o viceversa $F\,(N_A, N_B) = 1$. La medida de similitud cumple además con las propiedades de no negatividad, identidad, simetría, unicidad, desigualdad triangular y desigualdad triangular fuerte.

2.2.2.3 Similitud de Leacock-Chodorow

Con el objetivo de medir la similitud entre pares de palabras dentro del conjunto de datos *WordNet* (base de datos léxica del idioma inglés con sustantivos, verbos, adjetivos y adverbios), los autores (LEACOCK & CHODOROW, 1998) proponen la medida de similitud expuesta en la ecuación 2.7, donde D representa a la profundidad de la taxonomía (por ejemplo, en la figura 2.3, $D = 4$).

$$Sim_{lc}\,(N_A, N_B) = -\log\left(\frac{D1 + D2 + 1}{2 * D}\right) \tag{2.7}$$

2.2.2.4 Similitud de Resnik

Con el objetivo de evaluar la similitud semántica entre elementos representados en una taxonomía de red (lo cual es aplicable a problemas de inteligencia artifical, o psicología), se propone una medida que utiliza el concepto de C.I. descrito en la sección 2.2.2 a través de la ecuación 2.8 (RESNIK, 1995).

$$Sim_r\left(N_A, N_B\right) = -\log\left(p\left(A.C.M.\left(N_A, N_B\right)\right)\right) = C.I.\left(A.C.M.\left(N_A, N_B\right)\right) \tag{2.8}$$

2.2.2.5 Distancia de Jiang-Conrath

Luego de una revisión bibliográfica para analizar las ventajas y desventajas de los enfoques basados en nodos y aristas, el autor (JIANG & CONRATH, 1997) propone una medida de distancia semántica, descrita en la ecuación 2.9, aplicable al procesamiento del lenguaje natural, tratando las características de sinonimia y polisemia (elementos que generan ambigüedad).

$$Dist_{jc}\left(N_A, N_B\right) = C.I.\left(N_A\right) + C.I.\left(N_B\right) - \left(2 \times Sim_r\left(N_A, N_B\right)\right) \tag{2.9}$$

2.2.2.6 Similitud de Lin

Con el objetivo de desarrollar una medida de similitud semántica que no quede sujeta a una aplicación o dominio específico, en (LIN, 1998) se define una medida de similitud basada en un conjunto de intuiciones y supuestos (descritos en el apéndice B.2.1), la cual se presenta en la ecuación 2.10.

$$Sim_{lin}\left(N_A, N_B\right) = \frac{2 \times Sim_r\left(N_A, N_B\right)}{CI\left(N_A\right) + CI\left(N_B\right)} \tag{2.10}$$

2.2.3 Ejemplo comparativo entre medidas de distancia semántica

Para validar la efectividad de las medidas de distancia expuestas en la sección 2.2.2, se presenta un ejemplo del cálculo de los valores resultantes de la similitud semántica existente entre algunos elementos modelados a través de una estructura de DAG (misma que es utilizada en GO para modelar el conocimiento biológico), a través de las medidas de similitud presentes de las subsecciones 2.2.2.1 a 2.2.2.6.

La estructura de DAG referencial para obtener la distancia entre los nodos, es la representada en la figura 2.3. Se debe tener presente que la medida de la subsección 2.2.2.5 es una medida de distancia y por tanto se aplica su análogo como medida de similitud que corresponde al inverso de la distancia, para que sea coherente su comparación con el resto de las medidas.

Los nodos a evaluar de la figura 2.3 son los nodos C con respecto a D, G con respecto a H y D con respecto a F. Dado que los valores necesarios para el cálculo de las medidas de similitud con el enfoque de las aristas son *D1*, *D2*, *D3*, λ y D, estos se presentan en la tabla 2.3, considerando que si lo que se mide es $Sim\left(N_1, N_2\right)$, *D1* corresponde a la distancia de N_1 al A.C.M. entre N_1 y N_2, *D2* es la distancia de N_2 al mismo A.C.M., y *D3* es la distancia de ese A.C.M. al nodo raíz. Referente a los valores de frecuencia, probabilidad y C.I. necesarios para el cálculo de las medidas de similitud del enfoque basado en nodos, estos se exponen en la tabla 2.4. Los valores de similitud calculados se exponen en la tabla 2.5.

Ya obtenidos los valores de similitud entre los nodos analizados, se observa un efecto de otorgar mayor similitud a nodos que, si bien tienen visualmente la misma distancia entre sí, se encuentran en niveles inferiores en el DAG. Otro efecto claro, es el que produce que entre dos

TABLA 2.3: Valores requeridos para el cálculo de la medida de similitud semántica entre los nodos *C-D*, *G-H* y *D-F* presentes en la estructura de DAG de la figura 2.3, a través del enfoque basado en aristas.

Elemento	Sim(C, D)	Sim(G, H)	Sim(D, F)
N1	1	1	2
N2	1	1	2
N3	1	2	0
λ	0	0	0
D	4	4	4

TABLA 2.4: Valores requeridos para el cálculo de la medida de similitud semántica entre los nodos *C-D*, *G-H* y *D-F* presentes en la estructura de DAG de la figura 2.3, a través del enfoque basado en nodos.

Nodos	Frecuencia	Probabilidad	Contenido Información
Raíz	9	1	0
A	5	0.556	0.255
C	1	0.111	0.954
D	3	0.333	0.477
F	1	0.111	0.954
G	1	0.111	0.954
H	1	0.111	0.954

nodos el A.C.M. sea un nodo cercano a la raíz (o la misma raíz, en este caso) lo cual hace que entre ellos la similitud alcance valores cercanos a cero. Algunos efectos interesantes observables del ejemplo realizado, son el caso de la medida de similitud de *Leacock-Chodorow*, la cual entrega el mismo valor a la similitud entre C y D, que entre G y H, de hecho si se analiza con mayor profundidad, esa similitud es estática para cualquier par de nodos cuya distancia entre ellos sea de dos aristas, y la ontología cuente con cuatro niveles de profundidad (es decir, la similitud entre A y B y entre E y F también tiene ese valor), ello sucede porque no se considera algún valor asociado al A.C.M. Otro hecho interesante es el cálculo de la similitud de *Jiang-Conrath* por entregar el mayor valor de similitud a nodos vecinos de un nivel superior (C con D) por sobre nodos vecinos de un nivel inferior y por ende más específicos (G con H), ello no ha de considerarse un efecto negativo, sino más bien la capacidad de representar características sobre la estructura de ejemplo que las otras medidas no fueron capaces de considerar, dado que

TABLA 2.5: Valores de similitud semántica utilizando las medidas de las secciones 2.2.2.1 a 2.2.2.6, para algunos nodos de la estructura presentada en la figura 2.3

Medida de similitud semántica	Sim(C, D)	Sim(G, H)	Sim(D, F)
Wu-Palmer	0.500	0.667	0.000
T.B.K.	1.000	2.000	0.000
Leacock-Chodorow	0.426	0.426	0.204
Resnik	0.255	0.477	0.000
Jiang-Conrath	1.086	1.048	0.699
Lin	0.357	0.500	0.000

también es relevante el hecho de que la medida de *Jiang-Conrath* entregue el mayor valor para la similitud entre los nodos D y F (siendo incluso uno de los valores más altos dentro de la tabla).

2.2.4 Índice de homogeneidad biológica

Con el objetivo de identificar si un conjunto de genes tienen perfiles funcionales similares, se hace uso del *Índice de Homogeneidad Biológica* (IHB, ecuación 2.11), el cual fue propuesto por (DATTA & DATTA, 2006) y modificado por (VERBANCK et al., 2013) para ser aplicado a un grupo de genes (C_g). El índice evalúa por cada grupo (de cardinalidad $\#(C_g)$) el conjunto de términos biológicos obtenidos de GO que poseen los genes en dicho grupo ($Term$), analizando si cada gen del grupo tiene o no cada término del conjunto total (variable binaria $TermG_{g,i}$ que toma el valor cero cuando el gen g no posee la anotación i, y el valor uno en caso contrario), el total de términos asociados al gen g en evaluación ($Term_g$), el total de genes asociados al término i en evaluación (Gen_i), y el número total de relaciones entre genes y términos dentro del grupo ($Gen \times Term$) para entregar un valor en el intervalo $[0, 1]$, donde valores cercanos a cero indican que los genes del grupo no tienen perfiles funcionales similares, y un valor cercano a uno que sí los tienen. Dado que no todos los genes tienen conjuntos de términos asociados a ellos, al índice se le es aplicado un factor de tasa que divide al total de genes que tienen

términos ($GenTerm$) por el total de genes (Gen). El cálculo del índice es aplicado a todos los términos asociados a un gen (considera al término y a todos sus ancestros) para reforzar la relación entre genes con pocos términos específicos.

$$IHB\left(C_g\right) = \left(1 - \sqrt{\frac{\sum_{g \in C_g} \left(\sum_{i=1}^{Term} \frac{\left(TermG_{g,i} - \frac{Term_g Gen_i}{Gen \times Term} \right)^2}{\frac{Term_g Gen_i}{Gen \times Term}} \right)}{Gen \times Term \left(\#\left(C_g\right) - 1 \right)}} \right) \times \frac{GenTerm}{Gen} \tag{2.11}$$

2.3 ALGORITMO DE AGRUPAMIENTO *MST-KNN*

Entre las tareas involucradas en el análisis de datos de experimentos de expresión génica, está la organización de grupos de genes con características similares (agrupamiento) proceso que corresponde a la asignación de los genes de un conjunto a diferentes grupos, de manera que los de un mismo grupo tengan características similares, y los de grupos diferentes características que los hagan distintos. Los algoritmos encargados de realizar el agrupamiento se pueden dividir (conceptualmente) en supervisados (agrupan de acuerdo a valores de referencia) y no supervisados (no hay conocimiento previo para clasificar los datos). El algoritmo de agrupamiento no supervisado *MST-kNN*, utiliza como base la combinación de dos grafos de proximidad: *Minimum Spanning Tree* (árbol de expansión mínima, o *MST*) presente en el anexo A.2.1, y *k Nearest Neighbors* (k vecinos más cercanos, o *kNN*) el cual se presenta con mayor profundidad en el anexo A.2.2 (INOSTROZA-PONTA, 2008). Un grafo de proximidad mantiene dos vértices conectados si la relación entre ellos es acorde a una definición de proximidad dada, por ejemplo, que el peso de una arista que une dos vértices sea menor a un umbral definido.

En términos generales, si $E = \{e_1, e_2, ..., e_n\}$ es un conjunto con n elementos al cual se

le aplica un algoritmo agrupamiento, se puede obtener una partición del conjunto de la forma $C(E) = \{C_1, C_2, ..., C_m\}$, generando m subconjuntos disjuntos de E. Dentro de las etapas de un algoritmo de agrupamiento se encuentra la representación de los datos, definición de la medida de similitud, agrupamiento, abstracción de datos y evaluación de la salida (JAIN et al., 1999). La representación de los datos a ser agrupados es relevante cuando se trabaja con datos de expresión génica, pues de modelarse como un grafo (destacando relaciones entre los genes) se hace compatible con un algoritmo que utilice como parámetro de entrada grafos de proximidad (como *MST*, o *kNN*). Considerar la representación de grafo $G(V, A, P)$ para n genes (INOSTROZA-PONTA, 2008), de manera que:

- $\boldsymbol{V}$: Conjunto de vértices del grafo, donde cada vértice representa un gen. $V = \{v_1, v_2, ..., v_n\}$, $|V| = n$
- $\boldsymbol{A}$: Conjunto de aristas del grafo las cuales conectan dos vértices. $A : \{a_{ij} \backslash (i, j) \in (1..n, 1..n) \wedge i \neq j\}$, $|A| = n * \frac{n-1}{2}$
- $\boldsymbol{P}$: Conjunto de pesos para cada arista. El peso p_{ij} corresponde a la distancia entre los vértices i y j, calculado al utilizar una medida de distancia. $P : \{p_{ij} \backslash p_{ij} = d_{ij}\}$

Combinar grafos de proximidad como *MST* con *kNN* entrega mejores resultados a que si se utilizaran de forma separada, por ello en (INOSTROZA-PONTA, 2008) se propone una extensión al desarrollo presentado por (GONZÁLEZ-BARRIOS & QUIROZ, 2003), bajo la forma de un algoritmo de agrupamiento no supervisado con baja intervención del usuario, donde la idea es intersectar los grafos de proximidad *MST* y *kNN* eliminando así las aristas de *MST* que no estén presentes en la estructura *kNN* y realizando el cálculo automático del valor de k, dependiente de la cantidad de elementos, para generar *kNN* según:

$$k = \min \{ \lfloor \ln(n) \rfloor ; \min k \ / \ Graph_{kNN} \ se \ mantenga \ conectado \} \tag{2.12}$$

El algoritmo *MST-kNN* se aplica recursivamente en cada grupo que se genere luego de la intersección de los grafos de proximidad, hasta que el número de grupos obtenidos tras la intersección sea uno (*MST* no se divida en más grupos). El algoritmo 2.1 representa la forma de calcular las componentes conexas a partir de la matriz de distancia entre genes, considerando además los algoritmos A.1 y A.2 presentes en los apéndices A.2.1 y A.2.2 respectivamente.

Algoritmo 2.1: Pseudocódigo de la implementación de *MST-kNN* a partir de una matriz de distancia.

Data: D: Matriz de distancia.
Output: $Graph_{grupo}$: Grafo de grupos con $c \geq 1$ componentes conectados.

```
Calcular G;
Calcular Graph_MST;
Calcular Graph_kNN; // en base a lo visto en la expresión 2.12
Graph_grupo = {V_grupo = V, A_grupo = A_MST ∩ A_kNN}
c = componentesConectados (Graph_grupo);
if c > 1 then
    Graph_grupo = ⋃_{i=1}^{c} MSTkNN (subMatriz (D, Graph^i_grupo));
end
retornarGraph_grupo;
```

El algoritmo recibe una matriz de distancia de, por ejemplo, n objetos para con ella generar el grafo $G(V, A, P)$ que representa a los datos con un vértice por objeto y una arista por cada par de objetos con un peso igual a la distancia entre ambos. Luego, a partir del grafo inicial se calculan sus correspondientes $Graph_{MST}$ y $Graph_{kNN}$, los cuales se someten al algoritmo *MST-kNN* que retorna un grafo $Graph_{grupo}$, con $c \geq 1$ componentes conectados $\left(Graph_{grupo} = Graph^1_{grupo} \cup ... \cup Graph^c_{grupo}\right)$ y donde $A_{grupo} = A_{MST} \cap A_{kNN}$. Si el número de componentes conectados $c > 1$, se aplica el algoritmo a cada $\left(C^i_{grupo}\right)$ hasta que $c = 1$. El resultado final se compone de la unión de cada sub-grafo $Graph_{grupo}$. El algoritmo presenta pocos parámetros que definir facilitando su uso, y no asume una única medida de distancia

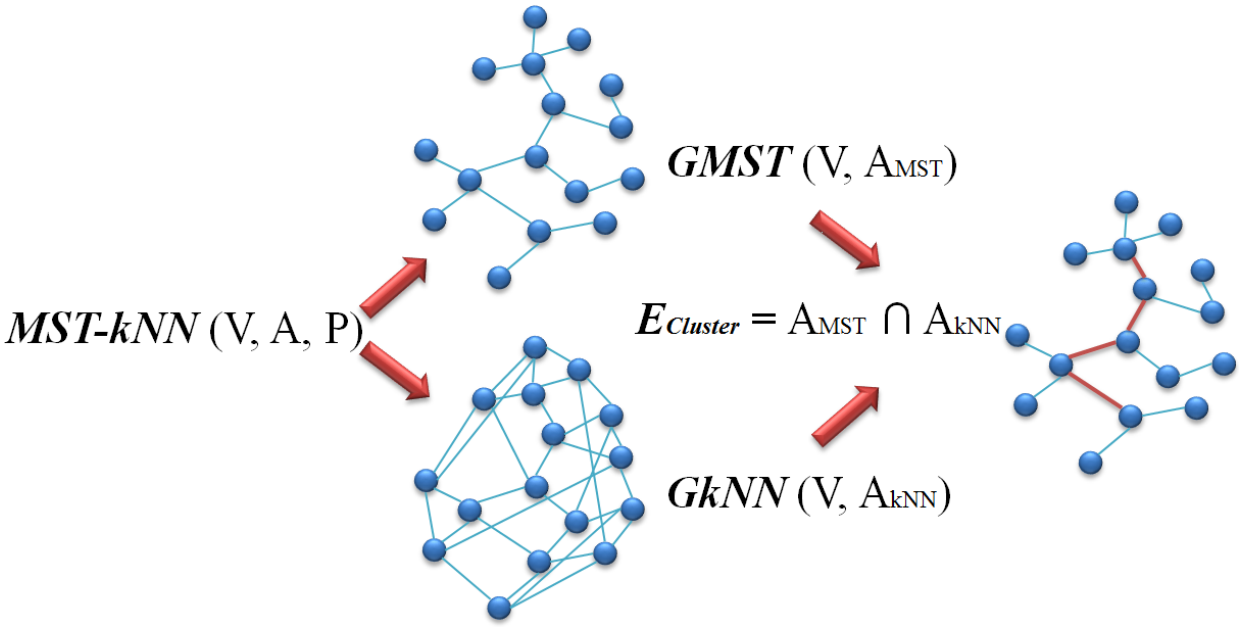

FIGURA 2.5: *Esquema del funcionamiento del algoritmo MST-kNN. A partir de una matriz de distancia se crea un grafo no dirigido $G(V, A, P)$, con un peso $p(a) \in P$ de una arista $a \in A$ representando una distancia entre dos vértices $v \in V$, que permite generar dos subgrafos $Graph_{MST}$, y $Graph_{kNN}$ que son posteriormente intersectados.*

para generar la matriz (y correspondiente grafo) lo que permite utilizarlo en variados escenarios (INOSTROZA-PONTA, 2008).

La figura 2.5 representa la generación del grafo $Graph_{MST}$ y el subgrafo $Graph_{kNN}$ a partir del grafo no dirigido $G(V, A, P)$, los cuales posteriormente son intersectados. En el grupo $E_{cluster}$ resultante, las líneas más delgadas de color cian representan a las aristas que se mantienen en la estructura final, mientras que las más gruesas, de color rojo, son las que se eliminaron particionando la estructura resultante y formando a los primeros grupos a los cuales se les aplica nuevamente el mismo procedimiento de manera separada.

Cabe mencionar que el algoritmo *MST-kNN* ha sido aplicado a diversos conjuntos de datos, como por ejemplo el relacionado a la esclerosis múltiple (DREW et al., 2010), o a genomas de mamíferos (CLARK et al., 2012).

2.4 DISCUSIÓN BIBLIOGRÁFICA

En términos generales, el problema a solucionar en la presente tesis se basa en cómo incorporar la información complementaria contenida en anotaciones biológicas, a los experimentos de expresión génica. Actualmente no hay muchos estudios que ofrezcan una alternativa de solución a dicho problema, sin embargo, a continuación se presentan los trabajos relacionados al desarrollo de estudios que implementen una incorporación de anotaciones biológicas a algoritmos de agrupamiento que utilizan datos de expresión génica, a nivel internacional.

El año 2010, se presentó una modificación del algoritmo *Super Paramagnetic Clustering* (CHERNOMORETZ, 2010) para generar grupos biológicamente significativos integrando la información de expresión génica de un experimento de *microarray*, con la información biológica de las funciones de los genes de la base de datos GO (sección 2.2.1). Como conclusión del trabajo, se obtuvo que añadir las anotaciones biológicas al proceso de agrupamiento mejoró la coherencia biológica de los grupos encontrados, en la identificación de procesos biológicos relevantes para los fenotipos de interés.

En otro trabajo, basándose en la premisa de que utilizar sólo datos de expresión para agrupar genes no se obtienen resultados adecuados en términos de correlación biológica, se propone un algoritmo de agrupamiento posibilístico semi-supervisado (*SemiSuperviced Possibilistic Clustering Algorithm*, SSPA) basado en restricción, integrando etiquetas o restricciones provistas por el usuario en un único proceso y no como una etapa posterior, lo que permite agrupar genes de acuerdo a su similitud (MARAZIOTIS et al., 2012). Las restricciones que guian el agrupamiento se obtienen de GO, y el método permite que un gen se asocie a más de una restricción pudiendo calcularse así el número de restricciones relacionadas a un gen y el número de violaciones a una restricción que posee un grupo, evaluando la calidad del agrupamiento. El algoritmo se comparó con PK-Medias, K-Medias y CPK-Medias utilizando

datos reales y artificiales, siendo superior a todos. El trabajo permite dar cuenta de los beneficios de la semi-supervisión aplicada a algoritmos de agrupamiento basados en expresión y la ventaja del uso de fuentes de información biológica externas como una guía para el agrupamiento de datos de expresión génica.

Otro trabajo propone un algoritmo de agrupamiento no supervisado que se basa en la integración del conocimiento biológico externo (anotaciones biológicas de GO) a los datos de expresión (VERBANCK et al., 2013). Para realizar dicha integración, proponen una nueva medida de distancia entre los genes, de manera que dos genes sean cercanos entre sí, si poseen perfiles de expresión similares y también perfiles funcionales similares. Además, proponen un procedimiento de evaluación para los grupos de genes con la ayuda de dos índices de validación, uno que permite medir la co-expresión global de los grupos, y otro que mide su homogeneidad biológica. La propuesta fue implementada y comparada con herramientas frecuentemente usadas en estudios de análisis de datos ómicos (mapas de calor y redes de co-expresión de genes), superándolos y entregando una proporción mayor de grupos de genes biológicamente homogéneos y significativamente co-expresados, los cuales son buenos candidatos para un posterior análisis por parte de biólogos para así formular nuevas hipótesis y establecer nuevas relaciones entre los genes de un organismo.

Con los conceptos vistos en este capítulo, debe quedar claro que se requiere de cuatro elementos fundamentales para realizar la incorporación de las anotaciones biológicas a los datos de expresión, para que éstos puedan ser agrupados por el algoritmo *MST-kNN*. Los elementos son: (1) una medida de distancia entre perfiles de expresión génica, (2) una medida de distancia entre términos biológicos de GO, (3) un índice de validación que mida la correlación de los perfiles de expresión de un grupo de genes, y (4) un índice de validación que mida la coherencia biológica de un grupo de genes. Cada uno de esos cuatro puntos, cuenta con variadas opciones posibles de ser implementadas, y de las cuales su aplicación varía de acuerdo a la calidad que presenten en un experimento determinado.

CAPÍTULO 3. INCORPORACIÓN DE ANOTACIONES BIOLÓGICAS

En el presente capítulo se describe el primer lineamiento de la solución propuesta a la pregunta fundamental planteada en la sección 1.2, que corresponde en términos generales al conjunto de pasos a seguir y funciones a utilizar para incorporar el conocimiento biológico a los datos de expresión génica, para así generar grupos de genes que contemplen ambas variables. El desarrollo de la solución se lleva a cabo en cinco etapas (representadas en la figura 3.1), donde para cada una de ellas hay un conjunto de métodos que pueden ser aplicados, y que han de ser coherentes con la consecución del objetivo general del trabajo (descrito en la sección 1.4.1).

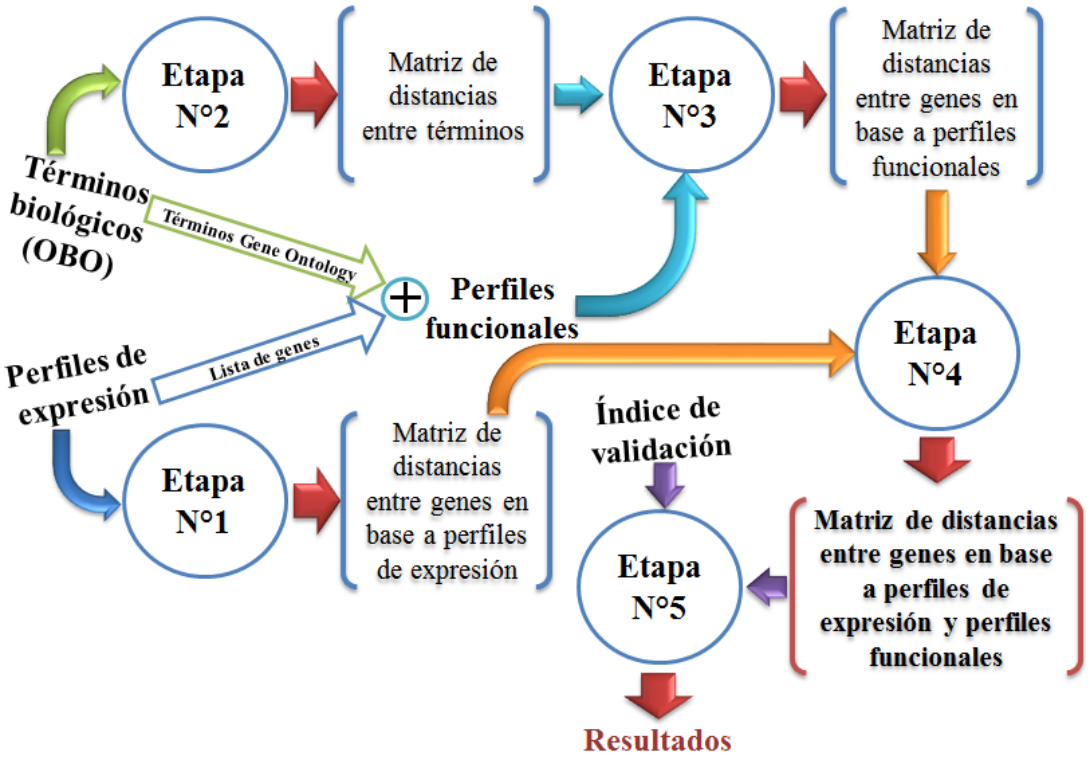

FIGURA 3.1: *Representación de las cinco etapas involucradas en la solución.*

Como se observa en la figura 3.1, la etapa número uno (descrita en la sección 3.1) establece una relación entre genes en base a sus perfiles de expresión. La segunda etapa, descrita en 3.2, permite establecer relaciones entre dos anotaciones biológicas de acuerdo a su ubicación dentro de una estructura de DAG (información provista por GO, o bien por el archivo OBO). Ya con el conjunto de genes participantes del estudio (los mismos utilizados para identificar sus perfiles de expresión), y la información provista por GO referente a los genes anotados en términos biológicos, se generan los perfiles funcionales de cada gen, para establecer las distancias entre genes en base a ellos (utilizando las distancias entre cada anotación biológica ya calculadas), proceso que es descrito en la sección 3.3. La etapa número cuatro (descrita en la sección 3.4), permite combinar la información de la relación entre genes en base a sus perfiles de expresión, con la relación entre genes en base a sus perfiles funcionales, obteniéndose una relación que incorpora ambos tipos de información y que es sometida al algoritmo de agrupamiento *MST-kNN* para evaluar, en una quinta etapa, los resultados obtenidos de la incorporación del conocimiento biológico a los datos de expresión génica.

3.1 DISTANCIA DE GENES SEGÚN SU EXPRESIÓN

Dado que el objetivo, según lo descrito en la sección 1.4.1 es agrupar genes por las similitudes que posean respecto a sus perfiles funcionales y perfiles de expresión, en la presente sección se describe como en una primera etapa los genes de la levadura *Saccharomyces cerevisiae* son agrupados de acuerdo a la similitud de sus expresiones génicas. Para ello, se genera una matriz de distancia entre genes de acuerdo a qué tan correlacionados están entre sí sus valores de expresión génica. Considerando la información descrita en la sección 2.1.1, se establece una correlación entre los perfiles de expresión de pares de genes (lo que permite calcular la distancia que existe entre ellos) de acuerdo a las ecuaciones siguientes:

$$Dist\,(g_A, g_B) = \frac{1 - \rho_{g_A,g_B}}{2} \tag{3.1}$$

$$Dist\,(g_A, g_B) = 1 - |\rho_{g_A,g_B}| \tag{3.2}$$

$$Dist\,(g_A, g_B) = \frac{2 - |(1 - (1 - \rho_{g_A,g_B})) + (1 - (1 - Spearman_{g_A,g_B}))|}{2} \tag{3.3}$$

Donde g_A y g_B representan a los perfiles de expresión de los genes A y B con los cuales se genera una matriz de distancia entre genes con valores en el intervalo $[0, 1]$, la cual puede someterse directamente al algoritmo de agrupamiento *MST-kNN* para establecer grupos de genes de acuerdo a sus perfiles de expresión.

3.2 DISTANCIA ENTRE TÉRMINOS POR SU SIMILITUD SEMÁNTICA

El proceso de relación entre genes en base al conocimiento biológico se lleva a cabo en dos etapas. La primera implica la generación de una matriz de distancia entre anotaciones biológicas, para en una segunda etapa generar una matriz de distancia entre genes considerando los conjuntos de anotaciones biológicas que los describen. Para la generación de la matriz de distancias entre anotaciones biológicas se hace uso de la información contenida en GO (sección 2.2.1), la cual al estar estructurada como un DAG, permite el uso de medidas de distancia semántica entre los nodos del mismo (sección 2.2.2). La obtención de las anotaciones biológicas necesarias se hace a través del acceso a la base de datos de GO, de manera que sólo se involucre a aquellos términos biológicos de la levadura *Saccharomyces cerevisiae* de la lista de genes utilizada en (EISEN et al., 1998), para luego establecer relaciones entre ellos (quién es ancestro o descendiente de quién) con ayuda de la información contenida en el archivo de OBO (apéndice

B.1.3). Así pues, los datos a utilizar se resumen como se indica a continuación:

```
Term_Name;Onto_Name;ID_Term;Rel;Term_Name;Onto_Name;ID_Term
Term_Name;Onto_Name;ID_Term;Rel;Term_Name;Onto_Name;ID_Term
...
```

Donde "*Term_Name*" indica el nombre de un término, "*Onto_Name*" la ontología a la que pertenece el término que le antecede, "*ID_Term*" el identificador único en la base de datos de GO del término en descripción, y "*Rel*" el tipo de relación que hay entre los dos términos, por ende en cada línea se describen dos términos (pertenecientes a una ontología específica) los cuales están relacionados según lo indique "*Rel*", lo que permite definir a un término como ancestro o descendiente de otro. Con las relaciones entre los términos resumidas se genera una estructura de DAG para realizar el cálculo de las distancias semánticas y con ello generar una matriz de distancias con valores en el intervalo $[0, 1]$ a través de una adaptación de las medidas descritas en las secciones 2.2.2.1 a 2.2.2.6 según las siguientes ecuaciones:

- Distancia de *Wu-Palmer*:

$$Dist_{wp} = 1 - Sim_{wp} \tag{3.4}$$

- Distancia de *T.B.K.*, donde D es la profundidad de la ontología:

$$Dist_{tbk} = 1 - \frac{Sim_{tbk}}{D} \tag{3.5}$$

- Distancia de *Leacock-Chodorow*:

$$Dist_{lc} = 1 - Sim_{lc} \tag{3.6}$$

- Distancia de *Resnik*: No se puede llevar a cabo la transformación de la medida de similitud, a una de distancia por tener valores en el intervalo $[0, \infty]$

- Distancia de *Jiang-Conrath*, donde T es el término más específico de la ontología:

$$Dist_{jc_{norm}} = \frac{Dist_{jc}}{2 \times C.I.(T)} \tag{3.7}$$

- Distancia de *Lin*:

$$Dist_{lin} = 1 - Sim_{lin} \tag{3.8}$$

3.3 DISTANCIA ENTRE GENES SEGÚN SUS PERFILES FUNCIONALES

Como se describe en la sección 3.2, la relación entre genes en base al conocimiento biológico se lleva a cabo en dos etapas. Ya resuelto el método para generar una matriz de distancia entre términos en base a su ubicación en la estructura de DAG de GO, se consideran los conjuntos de términos biológicos de la levadura *Saccharomyces cerevisiae* (1.032 términos biológicos) descritos según:

```
<Gene Annotation>
Gen_ID:Gen_Desc:Organism;GO_ID1;GO_ID2;...;GO_IDN
...
Gen_ID:Gen_Desc:Organism;GO_ID1;GO_ID2;...;GO_IDM
<EndOfFile>
```

Donde "*Gen_ID*" representa el identificador (o nombre) del gen, "*Gen_Desc*" la descripción del gen en cuestión, "*Organism*" el nombre del organismo al cual pertence el gen, y "*GO_IDi*" el i-ésimo identificador de un término de GO asociado al gen descrito. Con lo anterior se identifica al conjunto de términos asociados a un gen en particular, y con ello establecer las

distancias entre los conjuntos de términos de dos genes según:

$$GenTerm\left(g_i, g_j\right) = \begin{cases} 1 & g_i \vee g_j \; no \; tiene \; perfil \; funcional \\ Dist\left(g_i, g_j\right) & O.C. \end{cases}$$

De manera que $GenTerm\left(g_i, g_j\right)$ se evalúe según las siguientes funciones:

1. El valor mínimo de las distancias entre todas las anotaciones biológicas de los conjuntos de los genes. Dado que cada gen posee un conjunto de anotaciones biológicas que lo describe, se calculan todos los pares de distancias entre todas las anotaciones biológicas de ambos conjuntos y se considera el valor mínimo, como el valor representantivo de la distancia existente entre ambos genes.

2. El valor máximo de las distancias entre todas las anotaciones biológicas de los conjuntos de los genes. Al igual que con el valor mínimo, se calculan todos los pares de distancias de los conjuntos de anotaciones biológicas de los genes en cuestión, considerando al valor máximo como el representativo de la distancia existente entre ellos.

3. El promedio de las distancias entre todas las anotaciones biológicas de los conjuntos de los genes, es decir, se calculan todos los pares de distancias entre las anotaciones biológicas, y se considera al promedio de esas distancias como valor representativo de la distancia entre ellos.

4. El valor de la tasa entre todas las anotaciones biológicas de los conjuntos de los genes. Se calcula $\frac{T_{g_A} \cap T_{g_B}}{T_{g_A} \cup T_{g_B}}$, donde T_{g_i} es el conjunto de anotaciones biológicas de un gen i. Como utiliza los conceptos de unión e intersección de los conjuntos, el cálculo no hace uso de las medidas de distancia semántica descritas en la sección 3.2.

5. Calcular el promedio de las distancias ponderadas, es decir, se calculan todos los pares de distancia entre las anotaciones biológicas de los conjuntos de los genes y se ponderan

de acuerdo a la pertenencia de las anotaciones a los conjuntos según la ecuación 3.9. Este método permite utilizar dos criterios de evaluación que son el *matching* para cuando los elementos pertenecen a ambos conjuntos, y el promedio cuando no es así.

$$Dist_{PDP}\left(g_A, g_B\right) = \frac{1}{card\left(C_{g_A} \cup C_{g_B}\right)} \sum_{x \in \left(C_{g_A} \cup C_{g_B}\right)} f\left(x\right) \tag{3.9}$$

$$\text{Donde: } f\left(x\right) = \begin{cases} 0 & x \in \left(C_{g_A} \cap C_{g_B}\right) \\ Prom\left(dist\left(x, C_{g_B}\right)\right) & x \in C_{g_A} \wedge x \notin C_{g_B} \\ Prom\left(dist\left(x, C_{g_A}\right)\right) & x \notin C_{g_A} \wedge x \in C_{g_B} \end{cases}$$

6. Calcular el mínimo de las distancias ponderadas, donde al igual que el promedio de las distancias ponderadas se utilizan dos criterios. Se calculan todos los pares de distancia entre las anotaciones biológicas de los conjuntos de los genes y se ponderan de acuerdo a la pertenencia de las anotaciones a los conjuntos según la ecuación 3.10, considerando el concepto de *matching* para cuando los elementos pertenecen a ambos conjuntos, y el de valor mínimo cuando no se cumple esa condición.

$$Dist_{PDP}\left(g_A, g_B\right) = \frac{1}{card\left(C_{g_A} \cup C_{g_B}\right)} \sum_{x \in \left(C_{g_A} \cup C_{g_B}\right)} f\left(x\right) \tag{3.10}$$

$$\text{Donde: } f\left(x\right) = \begin{cases} 0 & x \in \left(C_{g_A} \cap C_{g_B}\right) \\ Min\left(dist\left(x, C_{g_B}\right)\right) & x \in C_{g_A} \wedge x \notin C_{g_B} \\ Min\left(dist\left(x, C_{g_A}\right)\right) & x \notin C_{g_A} \wedge x \in C_{g_B} \end{cases}$$

7. Calcular el *matching* entre todas las anotaciones biológicas de los conjuntos de los genes. Si se tiene dos genes g_1 y g_2, de manera que sus perfiles funcionales sean $T_{g_1} = \{t_1, t_2\}$ y $T_{g_2} = \{t_1, t_3\}$. El *matching* de la anotación t_1 del conjunto T_{g_1} con respecto a las anotaciones del conjunto T_{g_2} tiene una distancia de cero, pues dicho conjunto también tiene a la anotación t_1; el *matching* de la anotación t_2 del conjunto T_{g_1} con respecto a las anotaciones del conjunto T_{g_2} tiene un valor x_1 mínimo que puede obtenerse al relacionar a

la anotación t_2 con t_1, o con t_3, y por último el *matching* de la anotación t_3 del conjunto T_{g_2} con respecto a las anotaciones conjunto T_{g_1} tiene un valor x_2 mínimo que puede obtenerse al relacionar a la anotación t_3 con t_1, o con t_2. Luego, el cálculo del *matching* se obtiene al sumar cada valor de distancia calculada y dividirlo por el total de aristas, para el caso de ejemplo es: $\frac{0+x_1+x_2}{3}$

Con las siete opciones anteriormente descritas se obtiene un valor representativo de la distancia de los conjuntos de anotaciones biológicas que describen a un gen, y por ende, de la distancia entre dos genes que permite generar una matriz de distancia con valores en el intervalo $[0, 1]$.

3.4 DISTANCIA ENTRE GENES EN BASE A SUS PERFILES FUNCIONALES Y DE EXPRESIÓN

Dado que se tienen dos matrices de distancia entre genes, una en base a sus perfiles de expresión (sección 3.1) y otra a partir de sus perfiles funcionales (3.3), se vuelve a la pregunta que es el núcleo de la tesis: ¿de qué manera incorporar al agrupamiento de genes basado en datos de expresión, información de anotaciones biológicas de los genes?

Con el objetivo de no interferir directamente las características del algoritmo de agrupamiento descrito en (INOSTROZA-PONTA et al., 2011), se opta por utilizar el mismo parámetro de entrada de *MST-kNN* (sección 2.3), que es una matriz de distancias. El problema se reduce, entonces, a decidir una forma de combinar la matriz de distancia entre genes en base a perfiles de expresión, con la generada a partir de perfiles funcionales, para lo cual se hace uso de las siguientes opciones:

1. Utilizar un parámetro de ponderación α con intervalo de valores $[0, 1]$, de manera que un gen A tenga una distancia con un gen B ponderada según sus respectivas distancias

de expresión génica ($Dist_{expr_{g_A,g_B}}$) y de anotaciones biológicas ($Dist_{term_{g_A,g_B}}$), como lo indica la ecuación 3.11.

$$Dist_{alpha}\left(g_A, g_B\right) = \left[\alpha \times Dist_{expr_{g_A,g_B}}\right] + \left[(1-\alpha) \times Dist_{term_{g_A,g_B}}\right] \tag{3.11}$$

2. Utilizar la distancia euclídea entre ambos valores de distancia, considerando que cada una de ellas es una dimensión, con respecto al origen. Así pues, la distancia entre el gen A y el gen B con respecto a sus perfiles de expresión ($Dist_{expr_{g_A,g_B}}$) y perfiles funcionales ($Dist_{term_{g_A,g_B}}$), se define según lo descrito en la ecuación 3.12.

$$Dist_{eucl}\left(g_A, g_B\right) = \sqrt{Dist^2_{expr_{g_A,g_B}} + Dist^2_{term_{g_A,g_B}}} \tag{3.12}$$

Los elementos presentes en este capítulo describen las primeras cuatro, de las cinco etapas que conforman la base del desarrollo. Las cuatro primeras etapas describen la solución al problema planteado en la sección 1.2, y se resumen en: (1) relacionar genes de acuerdo a su comportamiento bajo las mismas condiciones experimentales, (2) relacionar anotaciones biológicas de acuerdo a su ubicación dentro de la estructura de DAG de la base de datos GO, (3) relacionar genes de acuerdo a las funciones a las cuales se asocian (conjunto de anotaciones biológicas), y (4) combinar la información de las relaciones entre genes de las etapas uno y tres, para conformar una única estructura que puede someterse al algoritmo de agrupamiento *MST-kNN* para generar grupos de genes similares de acuerdo a las características de su coexpresión, y de su coherencia biológica. Los grupos de genes son posteriormente evaluados en una quinta etapa, permitiendo definir los experimentos necesarios de llevar a cabo para validar la totalidad del desarrollo, proceso que se describe en el capítulo siguiente.

CAPÍTULO 4. PRUEBAS Y ANÁLISIS DE RESULTADOS

El presente capítulo tiene por objetivo describir tanto los experimentos realizados, como los resultados obtenidos a través de un análisis de los elementos fundamentales involucrados en el desarrollo, de manera que se establezcan de forma natural las conclusiones de la totalidad del trabajo.

4.1 DESCRIPCIÓN DE LOS EXPERIMENTOS

4.1.1 Metodología de evaluación

Como se observa en el capítulo tres, hay una serie de funciones y ecuaciones que combinadas generan como resultado la incorporación de conocimiento biológico al algoritmo de agrupamiento *MST-kNN*. Las etapas para llevar a cabo la incorporación de las anotaciones biológicas (figura 3.1), se describen a continuación:

I Generar una matriz de distancia entre genes en base a sus perfiles de expresión génica ($\mathcal{M}_{expr}$).

II Generar una matriz de distancia entre anotaciones biológicas en base a su ubicación dentro del DAG de GO ($\mathcal{M}_{term}$).

III Generar una matriz de distancia entre genes en base a sus perfiles funcionales ($\mathcal{M}_{func}$).

IV Combinar la matriz $\mathcal{M}_{expr}$ y $\mathcal{M}_{func}$, para obtener una matriz de distancia entre genes que incorpore ambos tipos de información ($\mathcal{M}_{inc}$).

V Evaluar los resultados obtenidos en el agrupamiento de genes (usando el algoritmo *MST-kNN*), cuando recibe como parámetro la matriz $\mathcal{M}_{inc}$, usando índices de validación que midan la correlación de perfiles de expresión, y la coherencia biológica de los genes dentro de los grupos generados.

Por cada una de las etapas anteriores hay una serie de opciones a considerar, y que generan diferentes parametrizaciones. Para la etapa I, se aplican las siguientes medidas de correlación (sección 3.1), todas normalizadas a un intervalo de valores entre $[0, 1]$:

1. Coeficiente de correlación de *Pearson* (notación: $1 - \rho$), a través de la ecuación 3.1.

2. Coeficiente de correlación de *Pearson* absoluto (notación: $1 - |\rho|$), a través de la ecuación 3.2.

3. Combinación del coeficiente de correlación de *Pearson*, con el de *Spearman* (notación: $\rho + \rho$), a través de la ecuación 3.3.

Para etapa II, se aplican las siguientes medidas de distancia semántica (sección 3.2), todas normalizadas a un intervalo de valores entre $[0, 1]$:

1. Medida de distancia que se basa en el conteo de aristas de (WU & PALMER, 1994), de acuerdo a la ecuación 3.4.

2. Medida de distancia basada en conteo de aristas de (SLIMANI et al., 2008), de acuerdo a la ecuación 3.5.

3. Medida de distancia que se basa en el conteo de aristas de (LEACOCK & CHODOROW, 1998), de acuerdo a la ecuación 3.6.

4. Medida de distancia que se basa en el análisis del contenido de información de nodos de (JIANG & CONRATH, 1997), de acuerdo a la ecuación 3.7.

5. Medida de distancia basada en el análisis del contenido de información de nodos de (LIN, 1998), de acuerdo a la ecuación 3.8.

Para la etapa III, se aplican las siguientes medidas de selección de distancias representativas, entre los conjuntos de anotaciones biológicas que poseen los genes (sección 3.3):

1. La distancia mínima entre los pares de distancias semánticas de los conjuntos de anotaciones biológicas de los genes (desde ahora, *Min*).

2. La distancia máxima entre los pares de distancias semánticas de los conjuntos de anotaciones biológicas de los genes (desde ahora, *Max*).

3. La distancia promedio entre los pares de distancias semánticas de los conjuntos de anotaciones biológicas de los genes (desde ahora, *Aver*).

4. La tasa entre los pares de distancias semánticas de los conjuntos de anotaciones biológicas de los genes (desde ahora, *Rate*).

5. El promedio de las distancias ponderadas, a través de la ecuación 3.9 (desde ahora, *PDP*).

6. El mínimo de las distancias ponderadas:, a través de la ecuación 3.10 (desde ahora, *MDP*).

7. El *matching* entre los conjuntos de anotaciones biológicas de los genes (desde ahora, *Match*).

Para la etapa IV, se aplican las dos funciones siguientes (sección 3.4):

1. Utilizar un ponderador α, según la ecuación 3.11, para un valor de $0,5$ (notación: $\alpha = 0,5$).

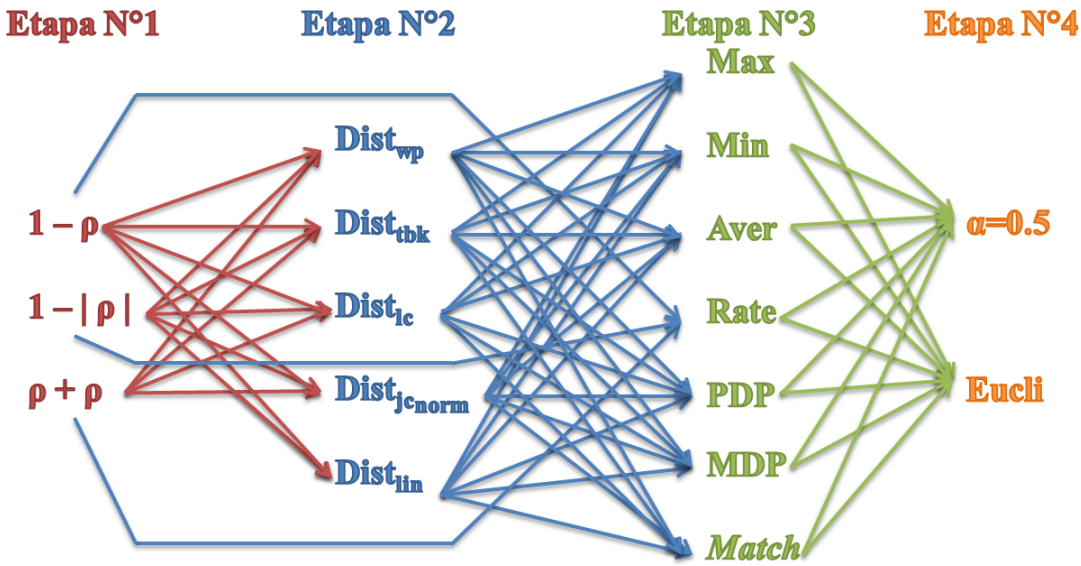

FIGURA 4.1: *Representación de las combinaciones posibles para generar matrices de distancia entre genes, utilizando las funciones y medidas de distancia implementadas, para posteriormente analizar los grupos que genere el algoritmo MST-kNN al recibir como parámetro de entrada dichas matrices.*

2. Utilizar la distancia euclídea, según la ecuación 3.12 (notación: *Eucli*).

Al considerar todas las combinaciones posibles de parámetros en cada uno de los cuatro puntos anteriores, se tiene un total de 186 posibles parametrizaciones (figura 4.1), las cuales generan distintas matrices $\mathcal{M}_{inc}$. La calidad del agrupamiento generado con cada matriz $\mathcal{M}_{inc}$ se evalúa con los siguientes índices:

1. Índice de coexpresión de los perfiles de expresión de los genes de un grupo (ecuación 2.2), con intervalo de valores $[-1, 1]$, donde -1 indica anti-correlación (los genes están sub-expresados), cero indica que no hay correlación y uno que hay correlación (los genes están sobre-expresados, notación: *IC*).

2. Índice de coexpresión absoluto de los perfiles de expresión de los genes de un grupo (ecuación 2.3), con intervalo de valores $[0, 1]$, donde cero indica que no hay correlación, y uno que hay correlación tanto si los genes del grupo están sobre-expresados como sub-expresados (notación: *IC Absoluto*).

3. Índice de homogeneidad biológica de un grupo (ecuación 2.11), con intervalo de valores $[0, 1]$, donde cero indica que los genes no comparten perfiles funcionales similares, y uno que los genes sí comparten perfiles funcionales similares. Para la evaluación de este índice se consideran sólo los genes descritos por un perfil funcional (tienen anotaciones biológicas que los describen, notación: *IHB*).

Así pues, los tres índices de validación anteriores se aplican a los grupos resultantes del agrupamiento generado por cada una de las 186 matrices $\mathcal{M}_{inc}$ obtenidas. Como los índices no se aplican al proceso de agrupamiento se hace necesario seleccionar valores de los índices que sean representativos de la parametrización completa, por lo que se utilizan dos medidas para representar la calidad de una parametrización, una que es en base al valor de los índices y otra que permite interpretar qué tan separados están los elementos de un grupo con respecto a los elementos de otros grupos. Ambas medidas se consideran funciones objetivo, por lo que se espera:

1. Maximizar el valor de un índice de validación para los grupos.

2. Maximizar la separación entre los grupos.

Respecto de la función objetivo número uno, y considerando a $Agr_{\mathcal{M}_{inc}}$ como el conjunto de grupos generado por una matriz $\mathcal{M}_{inc}$ para una parametrización, cada grupo del conjunto puede ser evaluado a través de los *IC*, *IC Absoluto*, o *IHB*. Ahora, si cada grupo tiene un valor determinado para algún índice surge la interrogante de, ¿cómo seleccionar un valor de índice representante de toda la parametrización? Para responderla se considera:

1. Calcular el promedio sólo de los grupos que tienen un valor de índice mayor a cero (afecta sólo a *IC*). Esta decisión evita el problema favorecer a grupos de alta cardinalidad y bajo valor de índice, o perjudicar a grupos con baja cardinalidad y alto valor de índice, y se justifica por el hecho conceptual de que no interesa someter a análisis a aquellos grupos cuyos genes no tienen correlación en sus perfiles de expresión.

2. Calcular el promedio de los grupos con valor de índice mayor que cero (dado el punto anterior), con el objetivo de maximizar el valor del índice en la representación de tuplas $(CardinalidadGrupo, ValorIndice)$, las cuales se pueden representar en un plano cartesiano bidimensional donde el eje de las ordenadas corresponda al valor de un índice, y el eje de las abscisas a la cardinalidad de los grupos. Dicho conjunto (de grupos representativos) recibirá el nombre de Agr_{max}. Tomando en consideración la razón de fondo del uso de la opción anterior, se promedian los valores de índices de aquellos grupos que representen a "los mejores" para una cardinalidad dada ($\overline{Agr}_{max}$). Considerar por ejemplo los valores de la tabla 4.1 donde la selección de los máximos se ve representada en la figura 4.2 por los puntos de color rojo, en ella se observa que dado un conjunto de grupos con un mismo valor de cardinalidad, se han de considerar los con mayor valor de índice para calcular el promedio, o valor representativo de la parametrización ($\overline{Agr}_{max} = 0,3875$ para los valores de la tabla 4.1). Se maximiza el valor del índice dado que los grupos pertenecientes al conjunto Agr_{max} representan a la información importante de analizar dentro de la parametrización, y ese razonamiento es aplicado con la misma prioridad para grupos con baja o alta cardinalidad, sin beneficiar a ninguno por sobre otro, pues no es una función objetivo la maximización o minimización de la cardinalidad de los grupos de una parametrización.

TABLA 4.1: Valores de ejemplo de tuplas $(ValorIndice, CardinalidadGrupo)$, para la selección de grupos que maximicen los valores de índice (función objetivo).

Grupo	Cardinalidad	Valor Índice	¿Es máximo?
A	2	0,1	No
B	2	0,3	Sí
C	3	0,45	Sí
D	3	0,35	No
E	4	0,1	No
F	4	0,25	Sí
G	5	0,55	Sí

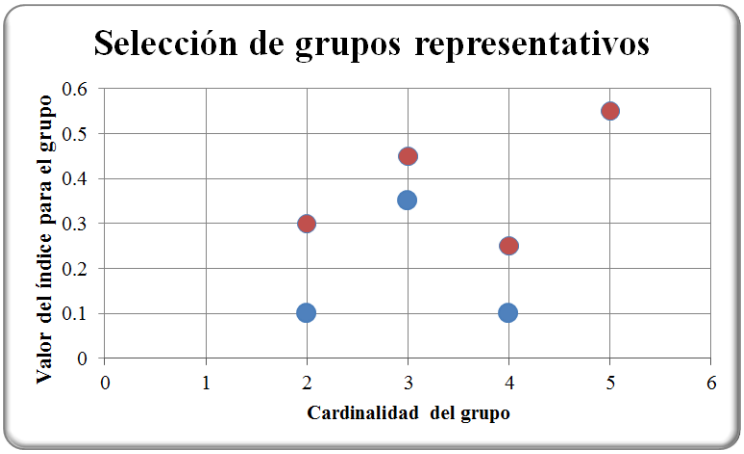

FIGURA 4.2: *Representación de valores expuestos en la tabla 4.1. En azul, los puntos que no son considerados para el valor promedio de la parametrización de acuerdo al índice, y en rojo los que sí pertenecen al conjunto que será promediado para obtener el valor representativo de la parametrización, por ser los valores máximos de cada cardinalidad (maximización de función objetivo).*

Respecto de la función objetivo número dos, se debe calcular un valor de separación entre grupos que sea representativo, para lo cual se opta por el promedio de las distancias promedio de los genes pertenecientes a un grupo, con respecto a todos los demás grupos en análisis (los que pertenecen al conjunto Agr_{max}). Considerar por ejemplo que luego de una parametrización, se generan N grupos, de los cuales sólo tres pertenecen al conjunto Agr_{max}: A, B y C (tal y como se describe en la figura 4.3), cada uno de cardinalidad dos. Para obtener la separación del grupo A con el grupo B (es decir, Sep_{A-B} con respecto a una medida de distancia), se promedian las distancias entre todo par (a, b), donde a y b son genes que cumplen con que $a \in A$, y $b \in B$; así pues, el promedio de las distancias promedio está dado por el cálculo de $Sep\,(param_i, dist_j) = \frac{Sep_{A-B}+Sep_{C-B}+Sep_{C-A}}{3}$, el cual es un valor representativo de qué tan separados están los grupos entre sí en la parametrización i, para la medida de distancia j. Se espera, por tanto, que dicho valor sea lo más alto posible para manifestar que los grupos generados están separados entre sí.

La medida de distancia a utilizar corresponderá a la misma usada en la parametrización, es decir, si en una parametrización A se utiliza $1 - |\rho|$ para generar la matriz $\mathcal{M}_{expr}$, entonces

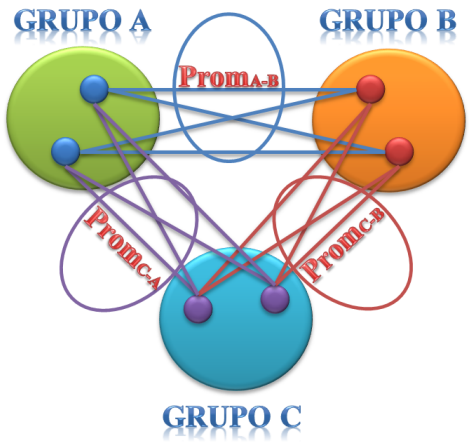

FIGURA 4.3: *Representación de valores del cálculo del promedio de las distancias para tres grupos de cardinalidad dos.*

se calcula la separación de los grupos generados con la matriz $\mathcal{M}_{inc}$ en base a $1 - |\rho|$. Todas las comparaciones son realizadas entre parametrizaciones y con respecto al agrupamiento entre genes en base a una matriz de distancia que utiliza sólo los perfiles de expresión, con el objetivo de identificar cómo varía la correlación de dichos perfiles, y la coherencia biológica de los perfiles funcionales con la incorporación de conocimiento biológico externo al proceso de agrupamiento.

FIGURA 4.4: *Plano cartesiano bidimensional de las parametrizaciones, con respecto al promedio de valores de un índice de validación y las distancias entre los elementos de los grupos. Los puntos en verde representan a parametrizaciones que son siempre inferiores a los que están en rojo, las cuales pertenecen al Conjunto de Pareto.*

Las medidas de distancia entre perfiles de expresión de genes forman parte de una variable importante de considerar a la hora de realizar un agrupamiento. Si el investigador desea, por ejemplo, encontrar relaciones entre genes considerando que una sub-expresión es

un comportamiento diferente a la sobre-expresión, entonces opta por utilizar una medida de distancia entre perfiles funcionales como $1-\rho$; si por otro lado desea evaluar si un gen reacciona bajo ciertas condiciones, independiente de si se sobre-expresa o sub-expresa, entonces opta por utilizar una medida de distancia entre los perfiles de expresión como $1-|\rho|$. Dadas las diferencias fundamentales de las medidas descritas, se llevan a cabo tres experimentos diferentes, es decir, de las 186 parametrizaciones posibles (representadas en la figura 4.1), se dividen de acuerdo a la medida de distancia de expresión génica ($1-\rho$, $1-|\rho|$ y $\rho+\rho$), más una parametrización que no considera la incorporación de anotaciones biológicas (es decir, para cada medida existen 63 parametrizaciones).

Para que la evaluación y análisis de los resultados abarque todos los tópicos de interés, se llevan a cabo siete preguntas, de manera que la respuesta a cada una de ellas guíe al reconocimiento de las conclusiones respectivas de la presente tesis. Las siete preguntas son:

1. Considerando que existen una serie de parametrizaciones posibles de seleccionar para trabajar con el conjunto de datos de la levadura *Saccharomyces cerevisiae*, ¿cuál o cuáles son las mejores?

2. Considerando los valores obtenidos para los índices de validación como un referente de cada parametrización, ¿se puede considerar la cantidad de grupos obtenidos como un indicador de la calidad esperada de una parametrización?

Las dos preguntas anteriores responden a una evaluación general de los tres experimentos descritos, permitiendo la obtención de las conclusiones principales, las cuales para ser corroboradas, se contrastan con la respuesta a las siguientes preguntas:

3. ¿Cuál de las tres medidas de distancia de expresión aparece con mayor frecuencia en las parametrizaciones mejor evaluadas?

4. ¿Cuál de los dos enfoques de medidas de distancia semántica aparecen con mayor frecuencia en las parametrizaciones mejor evaluadas?

5. ¿Cuál o cuáles medidas de similitud semántica aparecen con mayor frecuencia en las mejores parametrizaciones?

6. ¿Cuál o cuáles parámetros asociados a la selección de una distancia representativa de los conjuntos de anotaciones biológicas de los genes en estudio aparecen con mayor frecuencia en las mejores parametrizaciones?

7. ¿Cuál de las dos funciones para generar $\mathcal{M}_{inc}$ aparece con mayor frecuencia en las mejores parametrizaciones?

En las subsecciones siguientes se presentan las respuestas a cada una de las siete preguntas expuestas anteriormente, considerando la siguiente nomenclatura complementaria:

- *Separación matriz de expresión*: Valor promedio de la separación entre grupos generados por una parametrización i, en base a las distancias contenidas en la matriz $\mathcal{M}_{expr}$ de la misma parametrización i.

- *Separación matriz de términos*: Valor promedio de la separación entre grupos generados por una parametrización i, en base a las distancias contenidas en la matriz $\mathcal{M}_{func}$ de la misma parametrización i.

- $Experimento_{1-\rho}$: Experimento en que se utiliza a $1-\rho$ como medida de correlación entre los perfiles de expresión de los genes en estudio.

- $Experimento_{1-|\rho|}$: Experimento en que se utiliza a $1-|\rho|$ como medida de correlación entre los perfiles de expresión de los genes en estudio.

- $Experimento_{\rho+\rho}$: Experimento en que se utiliza a $\rho+\rho$ como medida de correlación entre los perfiles de expresión de los genes en estudio.

- D_{exp}: Distancia entre genes a partir de la similitud de sus perfiles de expresión génica.

- D_{term}: Distancia entre términos en base a su similitud semántica.
- $D_{exp-term}$: Distancia entre genes a partir de la similitud de sus perfiles funcionales.
- *Combinación*: Función de incorporación del conocimiento biológico a los datos de expresión génica.
- *S.I.*: Sigla de "Sin Incorporación", que representa a la parametrización que no utiliza conocimiento biológico externo para generar la matriz de distancia entre genes.

4.1.2 Datos de prueba

Se utilizó el conjunto de datos de la levadura *Saccharomyces cerevisiae* el cual tiene información de expresión génica de 2.467 genes con 79 muestras correspondientes a ocho experimentos: *alpha factor* (18 muestras), *cdc15* (15 muestras), *cold shock* (4 muestras), *diauxic shift* (7 muestras), *DTT shock* (4 muestras), *elutriation* (14 muestras), *heat shock* (6 muestras), y *sporulation* (11 muestras), todos provistos por Michael Eisen (EISEN et al., 1998). El conjunto de datos se mantiene en un archivo que describe a los de genes con sus valores de expresión génica, y los nombres de cada experimento:

```
<MicroarrayData>
2467,79
YBR166C,0.33,-0.17,0.04,...,-0.27
...
<SamplesNames>
alpha_0,alpha_7,alpha_14,...,diau_g
<EndOfFile>
```

Posteriormente se extraen las anotaciones biológicas de la base de datos GO de los mismos genes en estudio (correspondientes a la actualización del mes de Abril, año 2013), para luego hacer uso del archivo *OBO v1.2* (apéndice B.1.3) para obtener las relaciones entre los términos a los cuales se asocian los genes y formar con ellos la estructura de *DAG* correspondiente (donde toda relación utilizada en GO, se consdiera como una arista) que permite el cálculo de distancias

semánticas entre los conjuntos de términos de los genes (aplicando así las tres primeras etapas del proceso descrito en la figura 3.1).

A modo de resumen, la tabla 4.2 muestra el resumen de los conjuntos de datos que fueron utilizados para realizar las pruebas.

TABLA 4.2: Resumen del conjunto de datos utilizado en las pruebas de la solución implementada

#Genes	#Muestras de expresión (#experimentos)	#Términos	#Genes con términos
2.467	79 (8)	1.032	1.200

Tanto el desarrollo de la implementación de las cinco etapas relacionadas a la solución, como la ejecución de los experimentos y análisis de los resultados, se llevaron a cabo en un *Laptop* con S.O. *Windows 7* y *Ubuntu 12.04*, cuyas características de *Hardware* corresponden a una memoria RAM de 4GB, y un procesador *Intel Core i3* M350, de 2.27Ghz, arquitectura de 64 bits.

4.2 RESULTADOS DE EXPERIMENTOS

4.2.1 Mejor parametrización

Para definir la mejor parametrización se utilizan dos criterios: (1) el comportamiento de los genes con respecto al valor promedio de sus índices *IC*, *IC Absoluto* e *IHB* (de los grupos pertenecientes al conjunto Agr_{max} y según corresponda) y (2) el valor promedio de las separaciones de sus grupos ($Sep\,(param_i, dist_j)$), lo que puede generar más de una parametrización de buena calidad.

Para el experimento $Experimento_{1-\rho}$, se hacen tres análisis, todos con respecto a los índices de validación *IC* e *IHB*. El índice que mide la homogeneidad biológica es aplicable a cualquier tipo de experimento, pero *IC* es coherente aplicarlo sólo a $Experimento_{1-\rho}$, dado

que al utilizar $1 - \rho$ como distancia entre los perfiles de expresión tanto el índice como la medida de distancia buscan medir el mismo comportamiento en los genes. Esto es, que genes anticorrelacionados sean aquellos que presentan perfiles de expresión sub-expresados, y genes correlacionados los que tienen perfiles de expresión sobre-expresados. El primer análisis consiste en determinar al conjunto de mejores parametrizaciones de acuerdo a *IC* con respecto a la *Separación matriz de expresión*, para así identificar a aquellas parametrizaciones que generan grupos con la mayor correlación de sus perfiles de expresión, y mayor separación de acuerdo a $\mathcal{M}_{expr}$. El segundo análisis consiste en la determinación del conjunto de mejores parametrizaciones de acuerdo al *IHB* con respecto a *Separación matriz de términos*, con el objetivo de identificar a las parametrizaciones que generan grupos cuyos genes poseen mayor similitud en sus perfiles funcionales, y mayor separación de acuerdo a $\mathcal{M}_{func}$. El tercer análisis corresponde a la identificación de las mejores parametrizaciones de acuerdo a *IC* e *IHB* a la vez, es decir, aquellas que maximizan el valor de índices de validación y por tanto poseen los grupos cuyos genes tienen mayor correlación de sus perfiles de expresión y mayor similitud en sus perfiles funcionales.

TABLA 4.3: Parametrizaciones del conjunto de Pareto del $Experimento_{1-\rho}$, para la maximización de las funciones objetivo *IC* y *Separación matriz de expresión*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC*	*Separación*
S.I.	$1-\rho$	-	-	-	$0,049$	$0,417$
A	$1-\rho$	$Dist_{tbk}$	*PDP*	*Eucli*	$0,168$	$0,338$
B	$1-\rho$	$Dist_{jc_{norm}}$	*PDP*	$\alpha = 0,5$	$0,121$	$0,488$
C	$1-\rho$	$Dist_{wp}$	*Min*	$\alpha = 0,5$	$0,064$	$0,508$
D	$1-\rho$	$Dist_{lin}$	*Min*	$\alpha = 0,5$	$0,057$	$0,509$

Para el primer análisis, y según lo que se observa en la figura 4.5, las parametrizaciones que poseen los mejores valores para *IC* y además los más altos valores de *Separación matriz de expresión*, son cuatro (descritos en la tabla 4.3), las cuales se comparan además con los valores del agrupamiento de la parametrización *S.I.* Interesante es el hecho de que independiente del

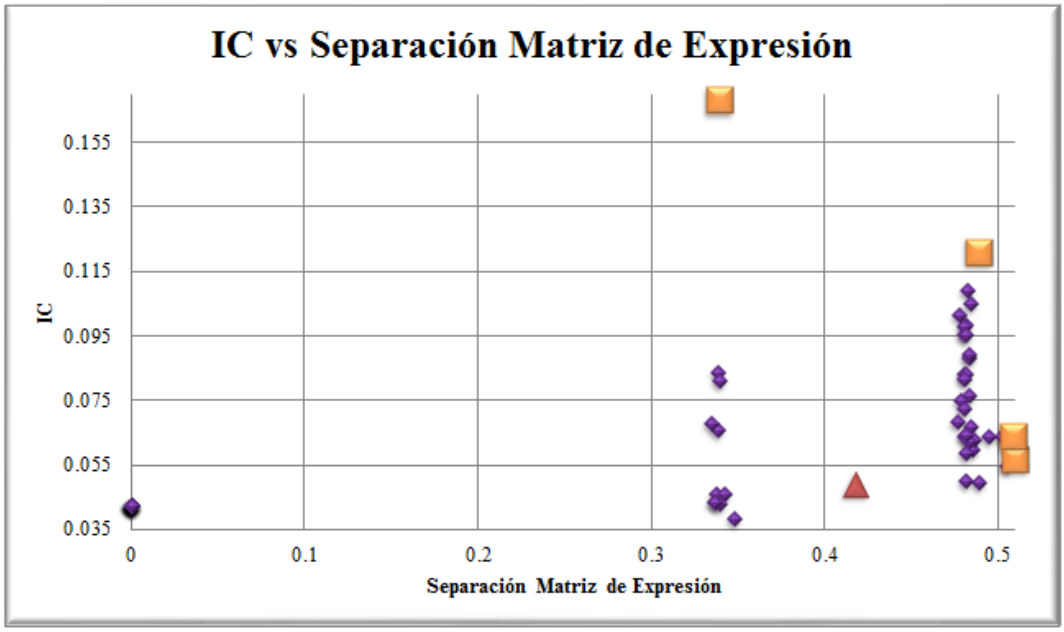

FIGURA 4.5: *Conjunto de Pareto del Experimento*$_{1-\rho}$*, para la maximización de las funciones objetivo IC y Separación matriz de expresión. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.3), y el triángulo de color rojo representa a la parametrización S.I.*

valor del índice, las separaciones se distribuyen en grupos o intervalos de valores cercanos, de acuerdo a la función que permite incorporar el conocimiento biológico. Comparativamente a no usar términos biológicos para generar el agrupamiento de genes, se tiene que un 59.68 % de las parametrizaciones superan en valor de *IC* al caso *S.I.*, donde la parametrización mejor evaluada posee una mejora porcentual del 242 % con respecto a *S.I.*. Respecto de las separaciones, se observa que un $50,00\,\%$ posee valores de separaciones más altos que *S.I.*, donde además la parametrización mejor evaluada con esa variable presenta una mejora porcentual de 22 %. Gráficamente, el triángulo rojo de la figura 4.5 representa a la parametrización *S.I.*, por tanto, todos los puntos que estén sobre él o a su derecha, son superiores a él ya sea con respecto a una variable, o con respecto a la otra.

Respecto del segundo análisis, dado que la parametrización *S.I.* no se le es posible de calcular un valor de *Separación matriz de términos* (pues no tiene una $\mathcal{M}_{func}$ asociada), se

le ha asignado un valor de distancia ficticio que corresponde al promedio de todos los valores de *Separación matriz de términos* de las 62 parametrizaciones posibles. Como se observa en la figura 4.6, existen nueve parametrizaciones en el conjunto de Pareto (descritas en la tabla 4.4), es decir, poseen los valores más altos de *IHB* y de *Separación matriz de términos* a la vez. Comparativamente a la parametrización que no incorpora anotaciones biológicas, se tiene que el 100 % de las parametrizaciones son superiores respecto del *IHB*, donde además la parametrización mejor evaluada posee una mejora porcentual de un 117 % con respecto ella. Por otro lado, existe un 48.39 % de parametrizaciones que tienen separaciones mayores (aunque en realidad sólo se indica que el 48.39 % de las parametrizaciones tienen valores superiores al promedio). Al analizar la figura 4.6, se denota un comportamiento variable donde se destaca una concentración de parametrizaciones con distancia entre grupos igual a cero (las cuales son las mismas presentes en el análisis anterior, observables en la figura 4.5), que representan a aquellas parametrizaciones que generaron sólo un grupo, lo que implica que no se puede calcular una distancia entre los genes de ese grupo y los demás (dado que no existen). El triángulo de color rojo representa a la parametrización *S.I.*, y se observa como todas son superiores a ella con respecto a *IHB*.

TABLA 4.4: Parametrizaciones del conjunto de Pareto del $Experimento_{1-\rho}$, para la maximización de las funciones objetivo *IHB* y *Separación matriz de términos*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IHB*	*Separación*
S.I.	$1-\rho$	-	-	-	$0,392$	$0,794$
A	$1-\rho$	$Dist_{lin}$	*PDP*	*Eucli*	$0,851$	$0,691$
B	$1-\rho$	$Dist_{tbk}$	*PDP*	*Eucli*	$0,825$	$0,699$
C	$1-\rho$	$Dist_{jc_{norm}}$	*PDP*	$\alpha = 0,5$	$0,783$	$0,888$
D	$1-\rho$	$Dist_{wp}$	*PDP*	$\alpha = 0,5$	$0,742$	$0,961$
E	$1-\rho$	$Dist_{lin}$	*PDP*	$\alpha = 0,5$	$0,730$	$0,965$
F	$1-\rho$	$Dist_{tbk}$	*Match*	$\alpha = 0,5$	$0,714$	$0,993$
G	$1-\rho$	$Dist_{tbk}$	*PDP*	$\alpha = 0,5$	$0,682$	$0,996$
H	$1-\rho$	$Dist_{tbk}$	*Aver*	$\alpha = 0,5$	$0,620$	$0,998$
I	$1-\rho$	-	*Rate*	$\alpha = 0,5$	$0,599$	$1,000$

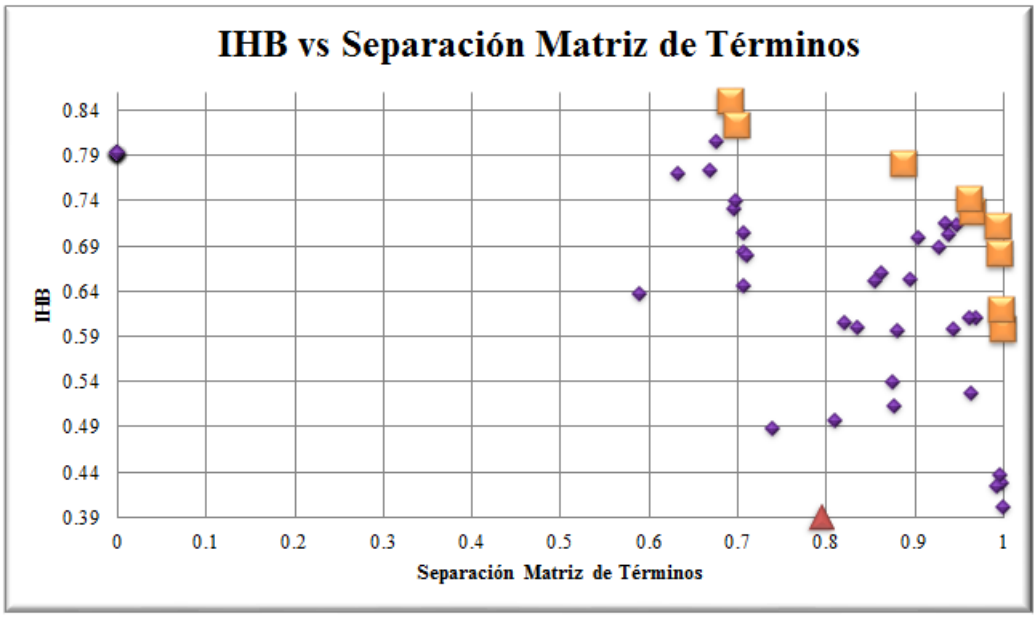

FIGURA 4.6: *Conjunto de Pareto del Experimento$_{1-\rho}$, para la maximización de las funciones objetivo IHB y Separación matriz de términos. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.4), y el triángulo de color rojo representa a la parametrización S.I.*

Para el último análisis del *Experimento*$_{1-\rho}$, se realiza la representación en un plano cartesiano bidimensional de las tuplas (IC, IHB), para así reconocer a las parametrizaciones que pertenecen al conjunto de Pareto, considerando que tanto el valor de *IC* como el de *IHB* deben ser los mayores (ambas funciones objetivo se maximizan). Como se observa en la figura 4.7 de las 62 parametrizaciones posibles hay dos que maximizan ambos valores de índices de validación, y corresponden a las parametrizaciones descritas en la tabla 4.5. Al comparar los valores de la totalidad de las parametrizaciones implementadas, con respecto a la parametrización *S.I.*, se tiene que un 59.68 % de ellas son superiores en valores de *IC*, y que el 100 % tienen un valor de *IHB* mayor. El comportamiento de las parametrizaciones se distribuye de manera dispersa entre ambas variables, pero es evidente la superioridad de las dos parametrizaciones pertenecientes al conjunto de Pareto, lo cual da cuenta de parametrizaciones adecuadas para el conjunto de datos, si se desea estudiar el comportamiento de los genes de acuerdo a un experimento como el de *Experimento*$_{1-\rho}$.

TABLA 4.5: Parametrizaciones del conjunto de Pareto del $Experimento_{1-\rho}$, para la maximización de las funciones objetivo *IC* e *IHB*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC*	*IHB*
S.I.	$1-\rho$	-	-	-	$0,049$	$0,392$
A	$1-\rho$	$Dist_{tbk}$	*PDP*	*Eucli*	$0,168$	$0,825$
B	$1-\rho$	$Dist_{lin}$	*PDP*	*Eucli*	$0,168$	$0,851$

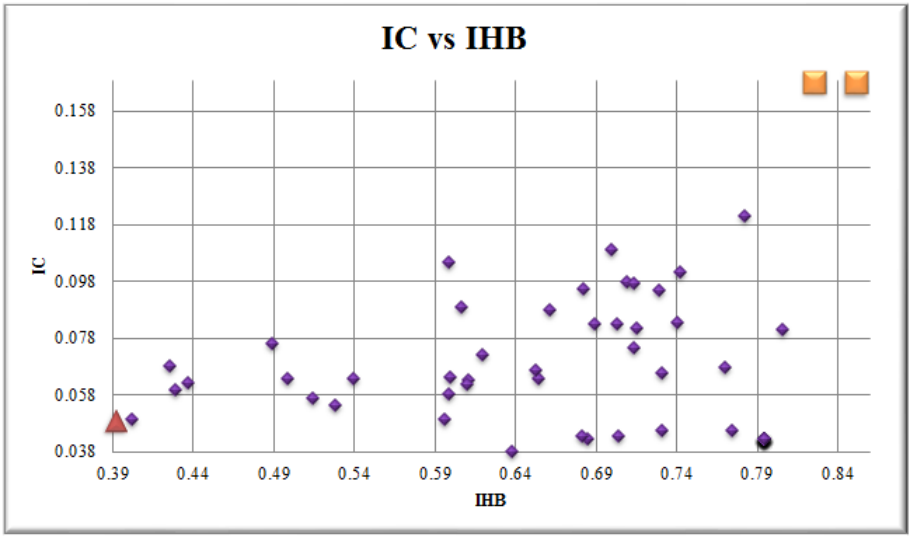

FIGURA 4.7: *Conjunto de Pareto del* $Experimento_{1-\rho}$*, para la maximización de las funciones objetivo IC e IHB. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.5), y el triángulo de color rojo representa a la parametrización S.I.*

Para el experimento $Experimento_{1-|\rho|}$, también se realizan tres análisis, pero esta vez con respecto a los índices de validación *IC Absoluto* e *IHB*. El índice que mide la homogeneidad biológica, como se mencionó anteriormente, es aplicable a cualquier tipo de experimento, pero *IC Absoluto* es coherente aplicarlo a $Experimento_{1-|\rho|}$ y $Experimento_{\rho+\rho}$, dado que al utilizar $1-|\rho|$ o $\rho+\rho$ como distancia entre los perfiles de expresión tanto el índice como la medida de distancia buscan medir el mismo comportamiento en los genes, es decir, que genes anticorrelacionados sean aquellos que presentan perfiles de expresión diferentes, y genes correlacionados los que tienen perfiles de expresión tanto sub-expresados como sobre-

expresados. El primer análisis consiste en determinar al conjunto de mejores parametrizaciones de acuerdo a *IC Absoluto* con respecto a la *Separación matriz de expresión*, para con ello identificar a las parametrizaciones que generan grupos con la mayor correlación de sus perfiles de expresión, y mayor separación de acuerdo a $\mathcal{M}_{expr}$. El segundo análisis consiste en la determinación del conjunto de mejores parametrizaciones de acuerdo al *IHB* con respecto a *Separación matriz de términos*, considerando nuevamente que para el caso de la parametrización *S.I.* se ha de utilizar el valor promedio de *Separación matriz de términos* obtenido en las 62 parametrizaciones implementadas en este experimento. Lo anterior tiene como objetivo identificar a las parametrizaciones que generan grupos cuyos genes poseen mayor similitud en sus perfiles funcionales, y mayor separación de acuerdo a $\mathcal{M}_{func}$. El tercer análisis corresponde a la identificación de las mejores parametrizaciones de acuerdo a *IC Absoluto* y al *IHB* al mismo tiempo, es decir, aquellas que maximizan los valores de los índices de validación.

Para el primer análisis, y según lo que se observa en la figura 4.8, las parametrizaciones con mayores valores para *IC Absoluto* y mayores valores de *Separación matriz de expresión* son seis (tabla 4.6), las cuales se comparan esta vez con los valores del agrupamiento de la parametrización *S.I.* (el cual en este caso utiliza la medida de distancia entre perfiles de expresión de $1 - |\rho|$). Interesante es el hecho de que independiente del valor del índice, las separaciones se distribuyen en grupos o intervalos de valores cercanos, de acuerdo a la función que permite incorporar el conocimiento biológico. Comparativamente al no uso de términos biológicos para generar el agrupamiento de genes, se tiene que un 37.10 % de las parametrizaciones superan en valor de *IC* al caso *S.I.*, donde la mejor parametrización evaluada presenta una mejora porcentual respecto a ella de un 4 %. Por otro lado, existe un $50,00$ % de parametrizaciones que poseen valores de separaciones más altos respecto de *S.I.*, donde la mejor evaluada presenta una mejora porcentual del 19 %. Gráficamente, el triángulo rojo de la figura 4.8 representa a la parametrización *S.I.*, por tanto, todos los puntos que estén sobre él son superiores a él con respecto al *IC Absoluto*, y los de su derecha son superiores con respecto a

la *Separación matriz de expresión*.

TABLA 4.6: Parametrizaciones del conjunto de Pareto del $Experimento_{1-|\rho|}$, para la maximización de las funciones objetivo *IC Absoluto* y *Separación matriz de expresión*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC Absoluto*	*Separación*
S.I.	$1-\|\rho\|$	-	-	-	$0,632$	$0,676$
A	$1-\|\rho\|$	$Dist_{tbk}$	*PDP*	$\alpha = 0,5$	$0,654$	$0,788$
B	$1-\|\rho\|$	-	*Rate*	$\alpha = 0,5$	$0,650$	$0,792$
C	$1-\|\rho\|$	$Dist_{tbk}$	*Min*	$\alpha = 0,5$	$0,634$	$0,798$
D	$1-\|\rho\|$	$Dist_{lin}$	*Min*	$\alpha = 0,5$	$0,627$	$0,800$
E	$1-\|\rho\|$	$Dist_{jc_{norm}}$	*Aver*	$\alpha = 0,5$	$0,625$	$0,803$
F	$1-\|\rho\|$	$Dist_{wp}$	*Min*	$\alpha = 0,5$	$0,622$	$0,804$

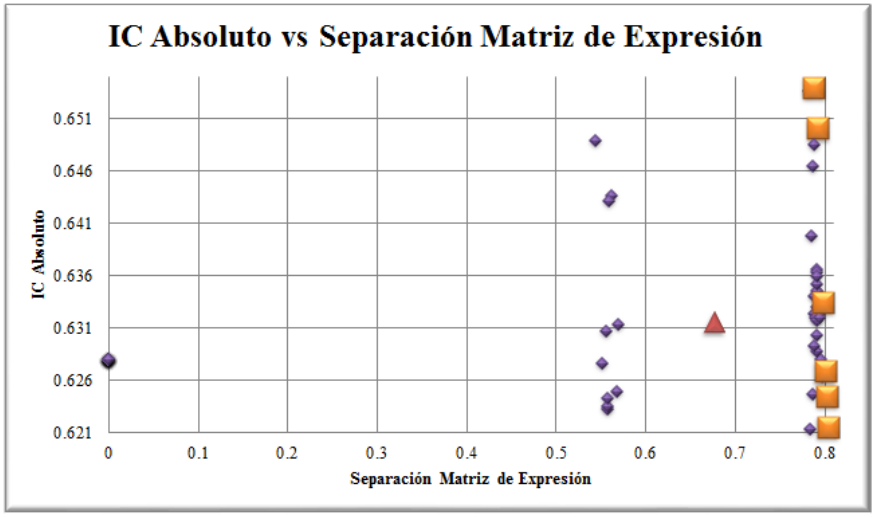

FIGURA 4.8: *Conjunto de Pareto del* $Experimento_{1-|\rho|}$*, para la maximización de las funciones objetivo IC Absoluto y Separación matriz de expresión. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.6), y el triángulo de color rojo representa a la parametrización S.I.*

Para el segundo análisis, en el cual nuevamente se considera como *Separación matriz de términos* para la parametrización *S.I.* al promedio de los valores de las 62 parametrizaciones implementadas, se observa en la figura 4.9 a siete parametrizaciones en el conjunto de Pareto, las cuales se describen en la tabla 4.7 y que generan grupos cuyos genes tienen los mayores valores de similitud de sus perfiles funcionales, y además las mayores separaciones con respecto

al conocimiento biológico. Comparativamente a la parametrización que no incorpora términos biológicos a los datos de expresión, se tiene nuevamente que el 100 % de las parametrizaciones implementadas poseen un valor de *IHB* mayor, donde la mejor evaluada incluso presenta una mejora porcentual del 113 %. Además, hay un 48.39 % de las parametrizaciones que tienen valores superiores al promedio de la *Separación matriz de términos* asignado al caso *S.I.*

TABLA 4.7: Parametrizaciones del conjunto de Pareto del $Experimento_{1-|\rho|}$, para la maximización de las funciones objetivo *IHB* y *Separación matriz de términos*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC Absoluto*	*Separación*
S.I.	$1-\rho$	-	-	-	$0,393$	$0,795$
A	$1-\|\rho\|$	$Dist_{jc_{norm}}$	*PDP*	*Eucli*	$0,836$	$0,639$
B	$1-\|\rho\|$	$Dist_{wp}$	*MDP*	*Eucli*	$0,753$	$0,655$
C	$1-\|\rho\|$	$Dist_{tbk}$	*PDP*	*Eucli*	$0,751$	$0,699$
D	$1-\|\rho\|$	$Dist_{lin}$	*PDP*	$\alpha = 0,5$	$0,724$	$0,966$
E	$1-\|\rho\|$	$Dist_{tbk}$	*MDP*	$\alpha = 0,5$	$0,708$	$0,993$
F	$1-\|\rho\|$	$Dist_{tbk}$	*PDP*	$\alpha = 0,5$	$0,707$	$0,995$
G	$1-\|\rho\|$	-	*Rate*	$\alpha = 0,5$	$0,596$	$1,000$

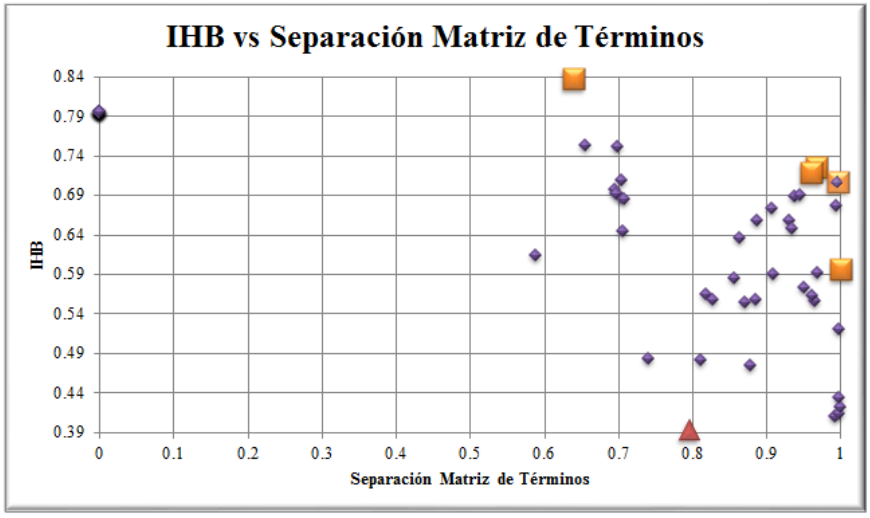

FIGURA 4.9: *Conjunto de Pareto del* $Experimento_{1-|\rho|}$*, para la maximización de las funciones objetivo IHB y Separación matriz de términos. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.7), y el triángulo de color rojo representa a la parametrización S.I.*

TABLA 4.8: Parametrizaciones del conjunto de Pareto del $Experimento_{1-|\rho|}$, para la maximización de las funciones objetivo *IC Absoluto* e *IHB*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC Absoluto*	*IHB*
S.I.	$1-\|\rho\|$	-	-	-	$0,632$	$0,393$
A	$1-\|\rho\|$	$Dist_{tbk}$	*MDP*	$\alpha = 0,5$	$0,654$	$0,708$
B	$1-\|\rho\|$	$Dist_{jc_{norm}}$	*PDP*	*Eucli*	$0,649$	$0,836$

El comporamiento de las parametrizaciones para los valores de las tuplas definidas es variable, destacándose nuevamente el conjunto de parametrizaciones con distancia entre grupos igual a cero, lo que representa a aquellas que generaron sólo un grupo. El triángulo rojo de la figura 4.9 (que representa a la parametrización *S.I.*) muestra claramente la superioridad de todas las demás con respecto al *IHB*.

El tercer análisis del experimento $Experimento_{1-|\rho|}$ da a conocer las parametrizaciones que poseen los mayores valores de *IC* e *IHB* a la vez, es decir, aquellas que son superiores al resto al maximizar la homogeneidad biológica y correlación de perfiles de expresión de los genes contenidos en los grupos que generan. La figura 4.10 expone que de las 63 parametrizaciones posibles sólo dos pertenecen al conjunto de Pareto (tabla 4.8), y además, al comparar los valores de la totalidad de las parametrizaciones implementadas con respecto a la parametrización *S.I.*, se identifica que el 35.48 % tiene valores mayores con respecto al *IC Absoluto*, y el 100 % tiene valores superiores con respecto al *IHB*. El comportamiento de las parametrizaciones se distribuye de forma dispersa entre ambas variables analizadas, lo que da cuenta de la evidente superioridad de todas las parametrizaciones con respecto al *IHB* (triángulo de color rojo en la figura 4.10). No es menor que, a diferencia del experimento $Experimento_{1-|\rho|}$, los elementos del conjunto de Pareto no sean extremadamente superiores a las demás parametrizaciones obteniéndose en general valores considerablemente altos para ambas variables en análisis.

Para el experimento $Experimento_{\rho+\rho}$, también se realizan tres análisis con respecto a los índices de validación *IC Absoluto* e *IHB*, dada su coherencia conceptual con lo que intenta medir la distancia entre perfiles de expresión de $\rho + \rho$. El primer análisis consiste en determinar al

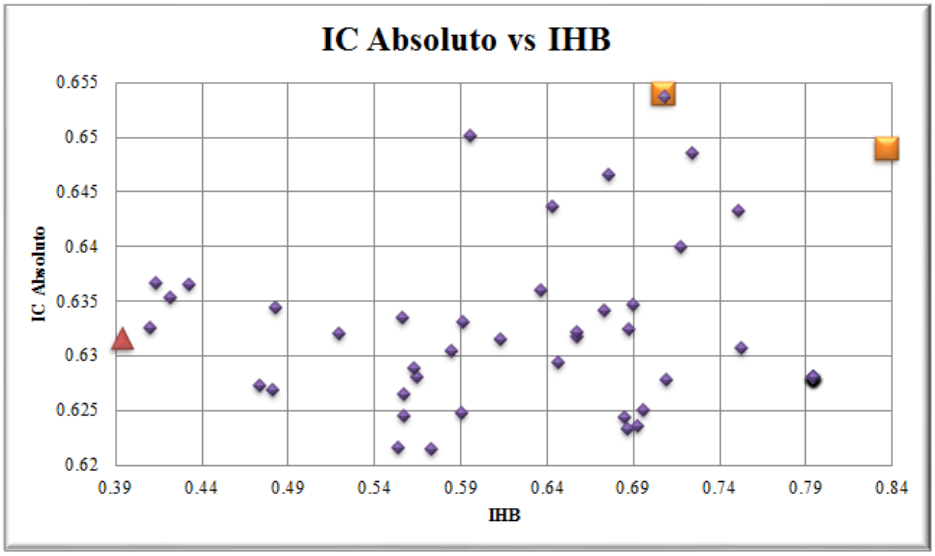

FIGURA 4.10: *Conjunto de Pareto del Experimento*$_{1-|\rho|}$*, para la maximización de las funciones objetivo IC Absoluto e IHB. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.8), y el triángulo de color rojo representa a la parametrización S.I.*

conjunto de mejores parametrizaciones de acuerdo al *IC Absoluto* con respecto a la *Separación matriz de expresión*, el segundo en la determinación del conjunto de mejores parametrizaciones para *IHB* con respecto a *Separación matriz de términos* (considerando para la parametrización *S.I.* el promedio de *Separación matriz de términos* de las 62 parametrizaciones implementadas), y el tercero a la identificación de las mejores parametrizaciones de acuerdo a *IC Absoluto* e *IHB* a la vez.

Respecto del primer análisis, se observan en la figura 4.11 las seis mejores parametrizaciones de acuerdo a la maximización de los valores del *IC Absoluto* y *Separación matriz de expresión* (descritos en la tabla 4.9). Nuevamente se detecta una distribución de distancias entre los genes de los grupos uniforme dependiendo de la selección entre las funciones $\alpha = 0,5$ y *Eucli*. Al comparar los valores de las parametrizaciones con el del caso *S.I.*, se tiene que un sólo un 12.90 % de las parametrizaciones poseen valores mayores de *IC Absoluto*, donde la mejor evaluada presenta una mejora porcentual del 1 % con respecto a ella. Por otro lado, se tiene un 50 % de parametrizaciones donde la *Separación matriz de expresión* sobrepasan al

caso *S.I.*, donde además la mejor clasificada posee una mejora porcentual del 18 %. Ese efecto se observa gráficamente, donde el triángulo de color rojo que representa a la parametrización *S.I.* posee una cantidad de puntos que están por sobre él (poseen mayor valor de *IC Absoluto*) que es menor que la cantidad de puntos que se ubican a su derecha (casos que poseen mayor *Separación matriz de expresión*).

TABLA 4.9: Parametrizaciones del conjunto de Pareto del $Experimento_{\rho+\rho}$, para la maximización de las funciones objetivo *IC Absoluto* y *Separación matriz de expresión*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC Absoluto*	*Separación*
S.I.	$\rho+\rho$	-	-	-	$0,641$	$0,694$
A	$\rho+\rho$	$Dist_{lin}$	*PDP*	$\alpha=0,5$	$0,649$	$0,806$
B	$\rho+\rho$	$Dist_{lin}$	*Max*	$\alpha=0,5$	$0,647$	$0,809$
C	$\rho+\rho$	$Dist_{lc}$	*Min*	$\alpha=0,5$	$0,643$	$0,810$
D	$\rho+\rho$	$Dist_{tbk}$	*Max*	$\alpha=0,5$	$0,641$	$0,812$
E	$\rho+\rho$	$Dist_{lc}$	*Max*	$\alpha=0,5$	$0,640$	$0,815$
F	$\rho+\rho$	$Dist_{jc_{norm}}$	*Aver*	$\alpha=0,5$	$0,633$	$0,819$

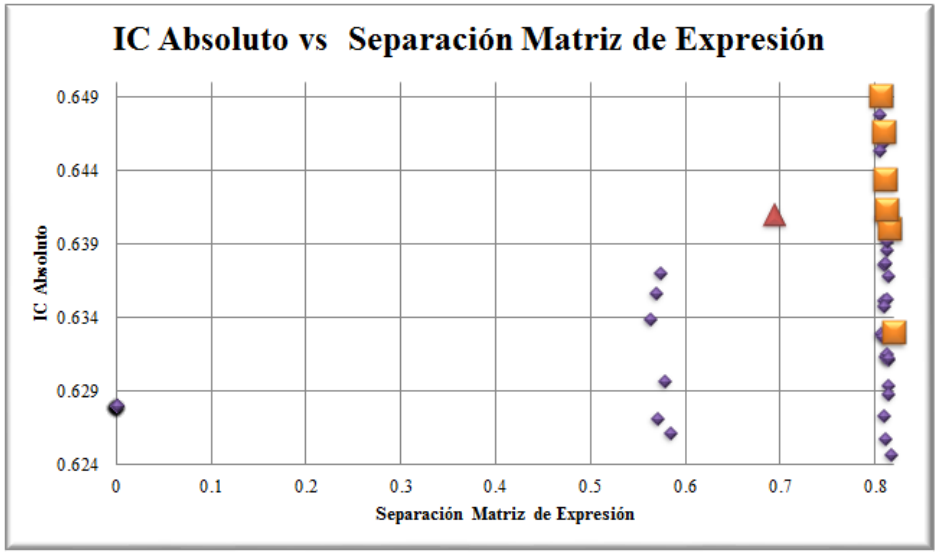

FIGURA 4.11: *Conjunto de Pareto del* $Experimento_{\rho+\rho}$*, para la maximización de las funciones objetivo IC Absoluto y Separación matriz de expresión. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.9), y el triángulo de color rojo representa a la parametrización S.I.*

Considerando como *Separación matriz de términos* de la parametrización *S.I.* al valor promedio de las 62 parametrizaciones implementadas, se observa en la figura 4.12 a las nueve parametrizaciones del conjunto de Pareto (descritas en la tabla 4.10) que, por tanto, generan grupos de genes con los mayores valores de similitud de sus perfiles de expresión y mayores separaciones de sus grupos con respecto al conocimiento biológico. Al comparar los resultados con la parametrización *S.I.*, se identifica que el 96.77 % de las parametrizaciones poseen un valor mayor de *IHB*, donde además la mejor evaluada presenta una mejora porcentual del 95 %. Por otro lado, se tiene un 49.39 % de parametrizaciones con una mayor separación de los grupos respecto de la información biológica (lo que en rigor indica que el 49.39 % posee un valor mayor al promedio). Referente al comportamiento de los datos, estos nuevamente se dividen en dos grupos, los que generan un agrupamiento con más de un grupo, y los que no. En este caso particular la parametrización *A* (de la tabla 4.10) es un valor representante de 26 parametrizaciones con el mismo valor de *IHB* y el mismo valor de *Separación matriz de términos* en el que se ven involucradas toda la gama de medidas de distancia entre términos biológicos, y todas las medidas de distancia entre genes en base a sus perfiles funcionales. Esas 26 parametrizaciones, comparten el mismo lugar por haber generado un único grupo luego de la aplicación del algoritmo *MST-kNN* a la matriz que generan, la cual tiene las características de: utilizar la medida de distancia entre perfiles de expresión de $\rho + \rho$ y utilizar la función *Eucli* para llevar a cabo la combinación de las matrices $\mathcal{M}_{expr}$ y $\mathcal{M}_{func}$ (y generar con ello a la matriz $\mathcal{M}_{inc}$). Lo anterior no es un hecho menor, dado que si bien el valor para del *IHB* no es bajo, no tiene sentido analizar a los 2.467 genes que quedaron en el único grupo formado. El criterio de Pareto, en este caso, basado en su forma conceptual a determinado que la parametrización *A* (y las otras 25) son buenas parametrizaciones, pero con un análisis mayor, se descarta de raíz la combinación de las funciones $\rho + \rho$ y *Eucli* para una misma parametrización.

TABLA 4.10: Parametrizaciones del conjunto de Pareto del $Experimento_{\rho+\rho}$, para la maximización de las funciones objetivo *IHB* y *Separación matriz de términos*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC Absoluto*	*Separación*
S.I.	$\rho+\rho$	-	-	-	$0,408$	$0,794$
A	$\rho+\rho$	$Dist_{lc}$	*Max*	*Eucli*	$0,795$	$0,000$
B	$\rho+\rho$	$Dist_{lc}$	*PDP*	*Eucli*	$0,750$	$0,656$
C	$\rho+\rho$	$Dist_{tbk}$	*Aver*	*Eucli*	$0,737$	$0,706$
D	$\rho+\rho$	$Dist_{lin}$	*Max*	*Eucli*	$0,729$	$0,710$
E	$\rho+\rho$	$Dist_{lin}$	*Match*	$\alpha=0,5$	$0,721$	$0,945$
F	$\rho+\rho$	$Dist_{lin}$	*PDP*	$\alpha=0,5$	$0,719$	$0,969$
G	$\rho+\rho$	$Dist_{tbk}$	*Match*	$\alpha=0,5$	$0,718$	$0,994$
H	$\rho+\rho$	$Dist_{tbk}$	*PDP*	$\alpha=0,5$	$0,709$	$0,996$
I	$\rho+\rho$	-	*Rate*	$\alpha=0,5$	$0,603$	$1,000$

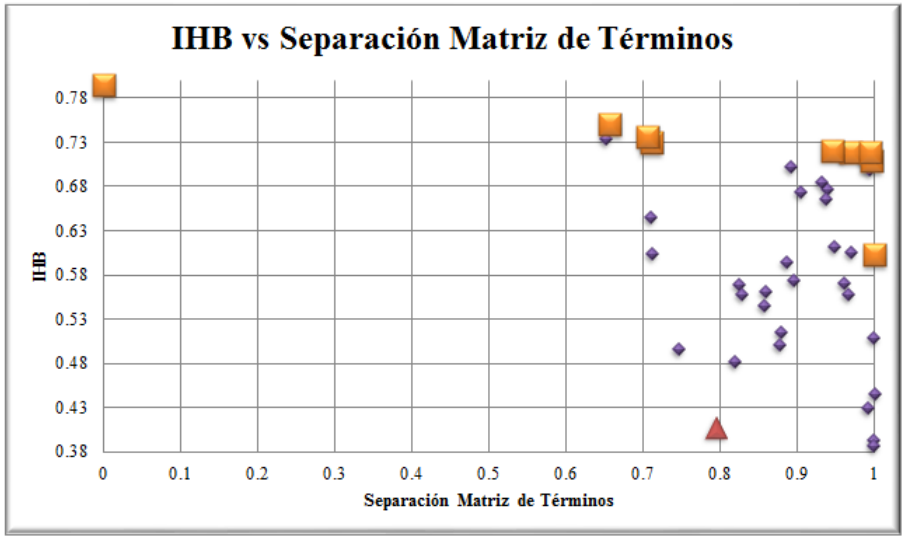

FIGURA 4.12: *Conjunto de Pareto del* $Experimento_{\rho+\rho}$*, para la maximización de las funciones objetivo IHB y Separación matriz de términos. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.10), y el triángulo de color rojo representa a la parametrización S.I.*

El tercer análisis relacionado a la selección de las parametrizaciones más adecuadas a partir de criterios definidos, hace relación con la selección de aquellas parametrizaciones del $Experimento_{\rho+\rho}$ que maximizan tanto el valor del *IC Absoluto* como el del *IHB*. La figura 4.13 representa a las 63 parametrizaciones, de las cuales sólo cinco pertenecen al conjunto de Pareto,

es decir, generan grupos cuyos valores de *IC Absoluto* e *IHB* son los mayores. Al comparar todas parametrizaciones implementadas con respecto a la parametrización *S.I.*, se observa que sólo el 12.30 % tiene valores mayores del *IC Absoluto*, mientras que el 96.77 % tiene valores superiores del *IHB*. En este caso, la figura 4.13 expone el hecho de que las parametrizaciones no tienen ningún comportamiento observable desde el punto de vista de los índices de validación, más allá del hecho de que sólo dos parametrizaciones no superan a la no incorporación de anotaciones biológicas, cuando se mide el *IHB*, de las cuales sorpresivamente se detecta que una de ellas, corresponde a la parametrización B de la tabla 4.9, lo que permite concluir que depende exclusivamente lo que se busque medir, los parámetros que se han de seleccionar, pues si se desea favorecer en una investigación específica los valores de expresión génica, se recomienda el uso de una parametrización como B de la tabla 4.9, pero si se desea estudiar la similitud funcional de genes, una parametrización como B de la tabla 4.9 es una opción completamente no recomendable.

TABLA 4.11: Parametrizaciones del conjunto de Pareto del $Experimento_{\rho+\rho}$, para la maximización de las funciones objetivo *IC Absoluto* e *IHB*, además de los valores de la parametrización *S.I.*

Caso	D_{exp}	D_{term}	$D_{exp-term}$	*Combinación*	*IC Absoluto*	*Separación*
S.I.	$\rho+\rho$	-	-	-	$0,641$	$0,408$
A	$\rho+\rho$	$Dist_{lin}$	*PDP*	$\alpha=0,5$	$0,649$	$0,719$
B	$\rho+\rho$	$Dist_{lin}$	*Match*	$\alpha=0,5$	$0,638$	$0,721$
C	$\rho+\rho$	$Dist_{tbk}$	*Aver*	*Eucli*	$0,637$	$0,737$
D	$\rho+\rho$	$Dist_{lc}$	*PDP*	*Eucli*	$0,636$	$0,750$
E	$\rho+\rho$	$Dist_{lc}$	*Max*	*Eucli*	$0,628$	$0,795$

Con toda la información expuesta y analizada anteriormente, se responde a la primera pregunta de análisis:

1. Considerando que existen una serie de funciones y ecuaciones posibles de seleccionar para trabajar con el conjunto de datos de la levadura *Saccharomyces cerevisiae*, ¿se puede seleccionar la parametrización más adecuada para dicho conjunto de datos? De ser posible

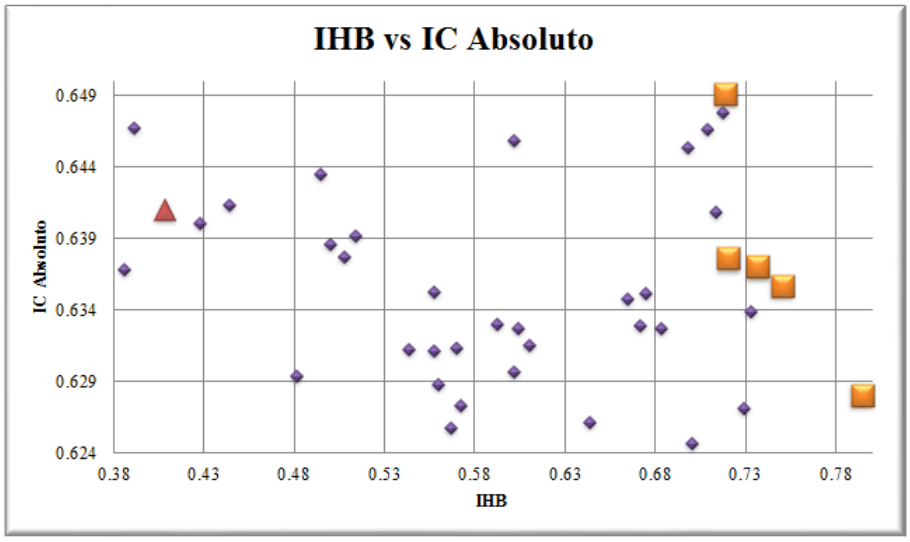

FIGURA 4.13: *Conjunto de Pareto del Experimento$_{\rho+\rho}$, para la maximización de las funciones objetivo IC Absoluto e IHB. Los cuadrados de color naranja representan a los que pertenecen al conjunto de Pareto (elementos descritos en la tabla 4.11), y el triángulo de color rojo representa a la parametrización S.I.*

¿cuál o cuáles son las mejores parametrizaciones o selecciones de funciones y ecuaciones?

Dado que la pregunta es amplia, se establecen alcances de qué es lo que generalmente se busca con la incorporación del conocimiento biológico a los experimentos enfocados en detectar similitudes entre genes. Como a nivel de agrupamiento se espera que los grupos que se generen sean homogeneamente coherentes respecto de los perfiles funcionales, y correlacionados respecto de los perfiles de expresión, se ha de considerar dicha variable como una función objetivo a maximizar. Dado que un agrupamiento, aparte de medirse qué tan similares son los elementos de los grupos entre si, tiene mejor o peor calidad de acuerdo a qué tan distintos son los elementos de un grupo con respecto a los elementos de otro grupo, utilizando como función objetivo la maximización de la distancia entre los grupos. Con esas dos variables, efectivamente es posible seleccionar a las mejores parametrizaciones, y de hecho, estas se encuentran expuestas en las tablas 4.3 a la 4.11. Esto hace evidente que la incorporación de anotaciones biológicas, genera grupos no sólo con una mayor homogeneidad biológica (con mejoras de hasta un 117 %), sino que además mejora la correlación de los perfiles de expresión de los grupos encontrados (con

mejoras de hasta un 242 %), lo que se traduce además en que los grupos posean una mayor separación entre sí.

La nueva variable incorporada al algoritmo de agrupamiento *MST-kNN*, efectivamente genera grupos de genes que son buenos candidatos a ser sometidos a un análisis mayor por parte de expertos en el área, pues éstos tienen una alta relación no sólo en base a su comportamiento en condiciones experimentales similares, sino que además comparten funciones biológicas, lo que permite, por ejemplo, determinar a un gen o grupo de genes involucrados en una enfermedad que antes no habían sido analizados, por no haber sido considerados como similares entre sí.

A pesar de los satisfactorios resultados presentados en esta sección, en que se corrobora el beneficio de la incorporación de conocimiento biológico al agrupamiento de genes, se ha de determinar el conjunto de variables y parámetros que son influyentes en la calidad de los resultados, para con ello entregar una gama de parametrizaciones que aseguren, de acuerdo a lo que se desee estudiar, resultados superiores a sólo utilizar perfiles de expresión para establecer similitudes entre genes.

4.2.2 Influencia del número de grupos

En el análisis de la sección 4.2.1, para el experimento $Experimento_{\rho+\rho}$, se consideró una parametrización perteneciente al conjunto de Pareto que generaba sólo un grupo, es decir, si se calcula la correlación entre todos los perfiles de expresión de todos los genes, se tiene un valor no menor para el *IHB*. Tras ese hecho surge la pregunta ¿hay alguna relación entre la cantidad de grupos y la calidad del agrupamiento? La presente sección responde a esa interrogante al analizar cómo se comportan los valores de los índices, a medida que la cantidad de grupos generados por las parametrizaciones aumenta o disminuye.

Se analizan nuevamente los tres experimentos ($Experimento_{1-\rho}$, $Experimento_{1-|\rho|}$ y $Experimento_{\rho+\rho}$) por separado, dado que buscan establecer una relación entre los genes bajo

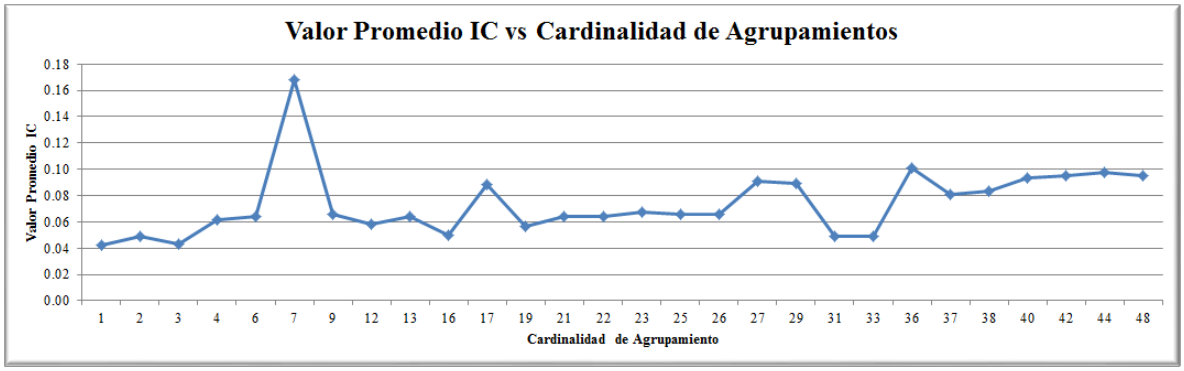

FIGURA 4.14: *Relación entre los valores del IC y la cantidad de grupos generados por el* $Experimento_{1-\rho}$.

criterios diferentes que no se han de mezclar entre sí. Las figuras 4.14 y 4.15 muestran el comportamiento del valor de los índices de validación *IC* e *IHB* respectivamente, con respecto a la cantidad de grupos generados por una parametrización dada para el $Experimento_{1-\rho}$. En ellas, se muestra que no hay un comportamiento o tendencia de los valores de los índices cuando aumenta o disminuye la cantidad de grupos generados, es decir, no importa si la cantidad de grupos es alta o baja, ello no refleja que la calidad de los mismos mejore o empeore. Ese hecho refleja además casos aislados, como el de cardinalidad del agrupamiento de valor siete con *IC* superior a $0,16$ (figura 4.14), el cual corresponde a la parametrización *A* de la tabla 4.3, es decir, una parametrización que se reconoce como de alta calidad. Interesante es el hecho de que, cuando se analiza la homogeneidad biológica de todos los genes de la levadura *Saccharomyces cerevisiae*, se encuentra un alto grado de relación (cercano a $0,80$ según la figura 4.15), lo cual da cuenta de algo más macro, pues indica que efectivamente las funciones biológicas asociadas a los genes tienen relación entre sí, y ello se debe, seguramente, a que pertenecen a una especie altamente estudiada. Ese valor de índice es, de hecho, uno de los más altos, pero al tener una separación nula entre sus grupos (pues existe un único grupo) no cumple con los requisitos necesarios para pertenecer al conjunto de Pareto, y es por ello que ninguna parametrización con ese índice aparece en la tabla 4.4.

Para el caso del experimento $Experimento_{1-|\rho|}$, en las figuras 4.16 y 4.17, que describen

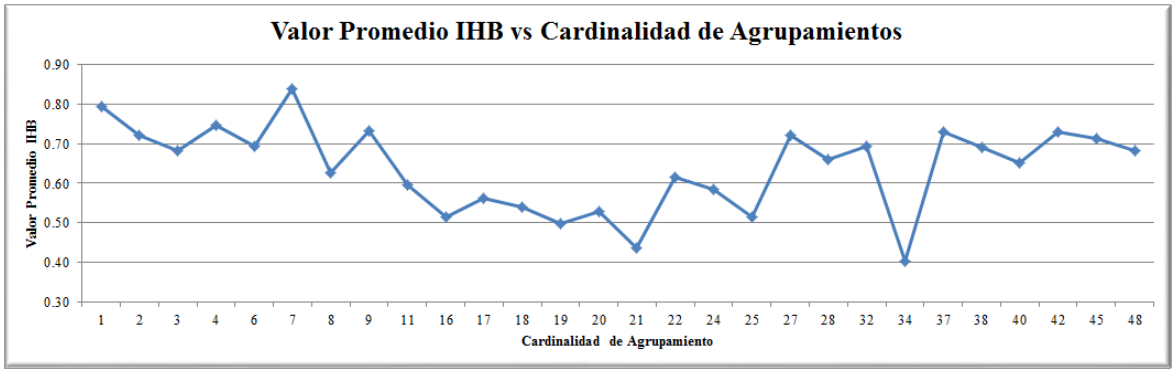

FIGURA 4.15: *Relación entre los valores del IHB y la cantidad de grupos generados por el* $Experimento_{1-\rho}$.

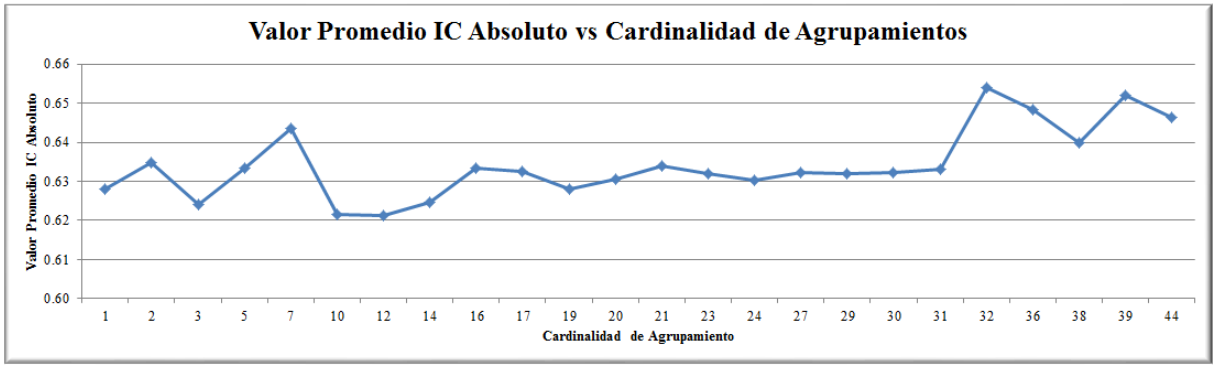

FIGURA 4.16: *Relación entre los valores del IC Absoluto y la cantidad de grupos generados por el* $Experimento_{1-|\rho|}$.

respectivamente el comportamiento de los valores de *IC Absoluto* e *IHB* con respecto a la cantidad de grupos generados por cada parametrización, tampoco se destaca un comportamiento creciente o decreciente a medida que aumenta o disminuye la cantidad de grupos que se generan, es decir, tampoco tiene relación alguna la cantidad de grupos con la calidad del agrupamiento. Para el caso de *IC Absoluto* se ve, de hecho, un comportamiento homogeneo en el intervalo $[15, 31]$ de la cantidad de grupos, y se tienen valores altos tanto si se generan siete grupos, como 44 grupos. A pesar de que el mismo efecto se da para el *IHB*, es interesante la alta calidad de homogeneidad biológica de los genes pertenecientes a grupos altamente poblados. El hecho anterior se debe a que tienen sentido las anotaciones biológicas asociadas a todo el conjunto de genes de la levadura *Saccharomyces cerevisiae*.

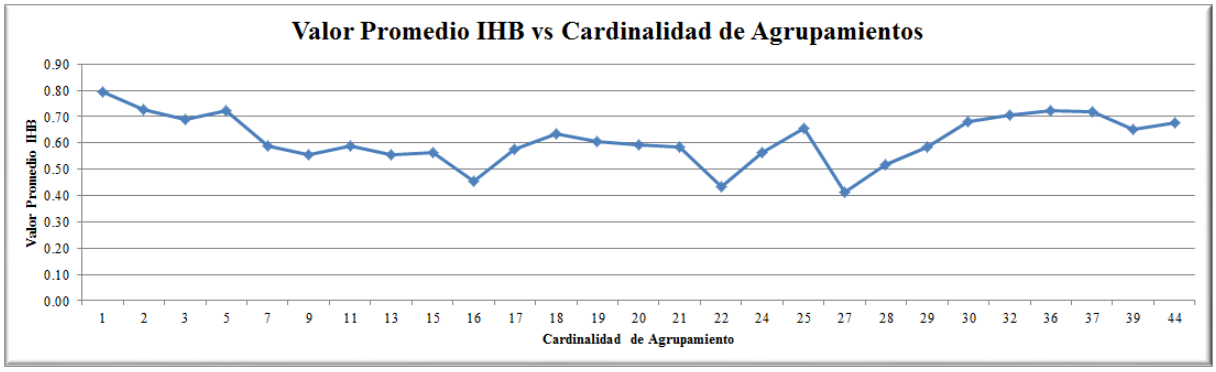

FIGURA 4.17: *Relación entre los valores del IHB y la cantidad de grupos generados por el* $Experimento_{1-|\rho|}$.

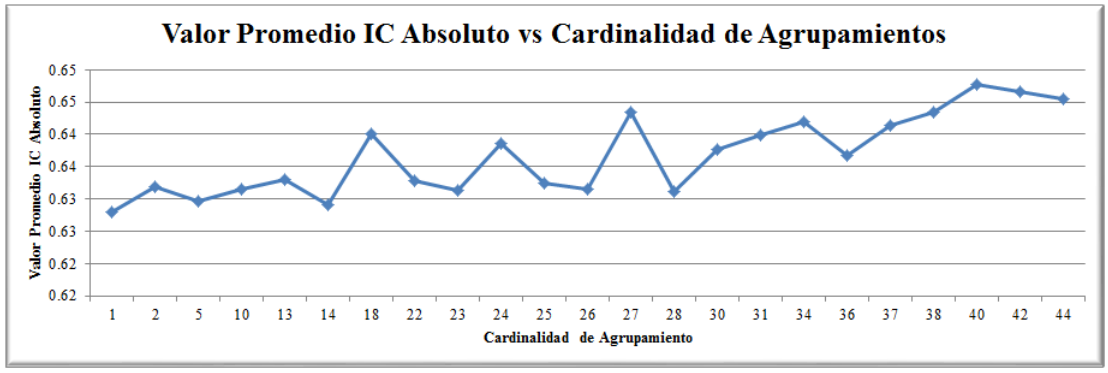

FIGURA 4.18: *Relación entre los valores del IC Absoluto y la cantidad de grupos generados por el* $Experimento_{\rho+\rho}$.

Por último, para el experimento $Experimento_{\rho+\rho}$, ocurre un caso similar a los anteriores, donde se observa en las figuras 4.18 y 4.19 que no hay sectores homogeneos para el *IC Absoluto* a media que aumenta o disminuyen la cantidad de grupos que generan las parametrizaciones. El valor más alto para *IC Absoluto*, lo tiene la parametrización A de la tabla 4.9, y ello es incluso con una alta cantidad de grupos generados a diferencia de, nuevamente, el alto valor del *IHB* para una parametrización que genera un único grupo. De lo anterior se extrae finalmente que no hay relación entre la cantidad de grupos que genere un agrupamiento, y la calidad de la parametrización lo cual es una conclusión relevante, pues se descarta de lleno utilizar la cantidad de grupos como una función objetivo que se ha de minizar o maximizar, debido a que no implica obtener resultados mejores o peores.

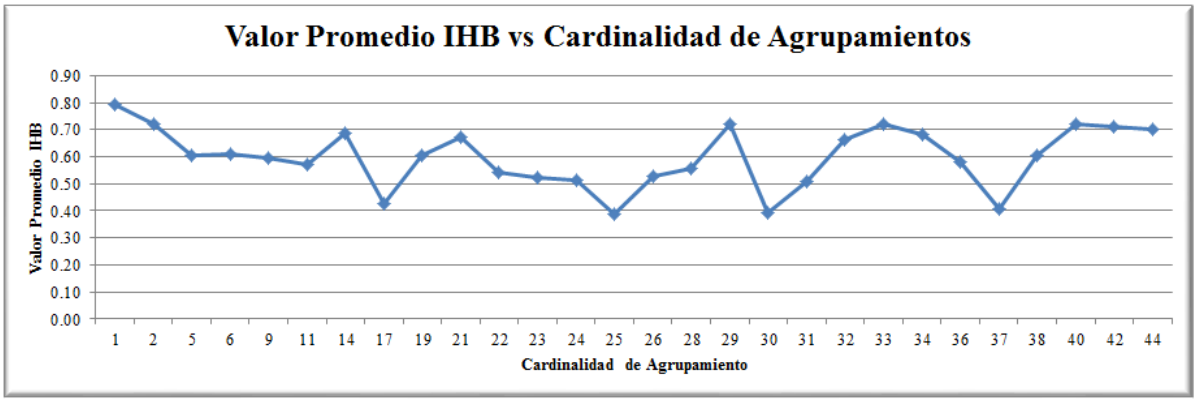

FIGURA 4.19: *Relación entre los valores del IHB y la cantidad de grupos generados por el* $Experimento_{\rho+\rho}$.

Dada la información anteriormente expuesta, se da respuesta a la segunda pregunta propuesta para el análisis de los resultados:

2. Considerando a los valores obtenidos para los índices de validación como un referente de cada parametrización, ¿se puede considerar a la variable "cantidad de grupos obtenidos" como influyente o representativa de una calidad esperada de una parametrización dada?

La respuesta es simplemente que *no*, la cantidad de grupos generados por una parametrización no es un índice de calidad válido para identificar si los genes de los grupos generados tienen relación entre sí. Claramente se ha de tener presente qué es lo que se busca, y esta variable es importante si es que el investigador lo considera así, por ejemplo, por muy bueno que sea el *IHB* para las parametrizaciones implementadas que generan sólo un grupo, no tiene sentido para él analizar 2.467 genes, porque de hecho, es esa la razón por la que se utiliza un algoritmo de agrupamiento, para agrupar a esos 2.467 en subgrupos coherentes e importantes de analizar. Pero lo anterior no es una verdad absoluta pues, por ejemplo, un investigador puede querer analizar si todas las anotaciones asignadas a una misma especie tienen coherencia entre sí, caso en que sí es necesario que la homogeneidad biológica de los 2.467 genes (para el caso de la levadura *Saccharomyces cerevisiae*) tenga un alto valor.

4.2.3 Mejor medida de distancia de perfiles de expresión génica

Las secciones anteriores tenían por objetivo entregar características generales de una parametrización completa, es decir, cómo la incorporación de información afecta a la calidad de los resultados. Ahora que se conocen las mejores combinaciones de parametrizaciones (sección 4.2.1), o la relación entre la cantidad de grupos generados y la calidad de una parametrización (sección 4.2.2), lo siguiente es analizar ¿cuál parámetro en cada una de las etapas entrega los mejores resultados? Ello con el objetivo de responder las cinco preguntas restantes planteadas al inicio del análisis.

La presente sección, como su nombre lo indica, pretende responder a la tercer interrogante propuesta en un inicio del análisis:

3. ¿Cuál de las tres medidas de distancia entre genes en base a sus perfiles de expresión se ve mayormente reflejada en las parametrizaciones mejor evaluadas?

Para responderla, basta con clasificar las 186 parametrizaciones con respecto a los valores promedio obtenidos para algún índice de validación, y de acuerdo al orden generado, identificar si $1-\rho$, $1-|\rho|$ o $\rho+\rho$ se ve mayormente presente en los 10 ó 5 casos mejor clasificados, pero ¿tiene sentido aquello? Numéricamente se puede lograr, pero no tiene sentido mezclar los tres experimentos ($Experimento_{1-\rho}$, $Experimento_{1-|\rho|}$ y $Experimento_{\rho+\rho}$) puesto que cada medida de distancia en base a la correlación de los perfiles de expresión de los genes es propuesta con un objetivo diferente, por tanto, no se puede esperar que $Experimento_{1-|\rho|}$ sea mejor que $Experimento_{1-\rho}$, si lo que se desea es que genes anticorrelacionados sean aquellos cuyos perfiles de expresión estén sub-expresados. Dado lo anterior, no se puede responder a la pregunta, puesto que al dividir por experimento, el 100 % de las parametrizaciones mejor evaluadas corresponderan a $1-\rho$, $1-|\rho|$ o $\rho+\rho$ según sea el caso. El investigador debe tener claro qué es lo que desea, para el menos este primer parámetro considerarlo de acuerdo a esa necesidad.

4.2.4 Mejor enfoque de distancia semántica

Con el objetivo de contestar la cuarta pregunta expuesta para el análisis de los resultados, se ve cada experimento por separado clasificando las 63 parametrizaciones de cada uno según el valor de un índice de validación acorde al experimento, para así identificar cómo se distribuyen los enfoques de similitud semántica entre términos biológicos.

La figura 4.20 representa al 50 % de las mejores clasificaciones obtenidas con respecto a los índices *IC* e *IHB* (en la figura C.1 se observa la totalidad de la clasificación). En la figura, las parametrizaciones se ordenan de forma decreciente, desde arriba hacia abajo, y se destaca el 5 %, 10 %, 25 % y 50 % mejor clasificado. Dado lo anterior, para el experimento $Experimento_{1-\rho}$ con respecto al *IC* se observa que es el enfoque basado en las aristas el que entrega los valores más altos, mientras que la clasificación número 38 corresponde a la parametrización *S.I.* En general, el enfoque basado en las aristas se ve mayormente representado que el enfoque basado en los nodos, pero ello se debe a que fueron utilizadas tres medidas de distancia entre términos basadas en las aristas, mientras que del enfoque basado en los nodos sólo fueron utilizadas dos. Interesante es el hecho que no utilizar un enfoque basado en distancias semánticas (*Rate*) entrega resultados desde la quinta clasificación, lo que da cuenta de que para medir la coexpresión entre genes, no es prioritario utilizar una medida que distancia entre los genes en base a sus conjuntos de términos que analice más allá de cuáles tienen en común y cuáles no. Es importante notar que los resultados coinciden con los expuestos en la tabla 4.3, pues el conjunto de Pareto está conformado por las medias de $Dist_{tbk}$, $Dist_{jc_{norm}}$, $Dist_{wp}$ y $Dist_{lin}$, es decir, dos basadas en los nodos y dos en las aristas.

Para el caso en que se clasifica respecto del *IHB*, se observa que se invierte el enfoque mejor evaluado, pues en ese caso es el basado en los nodos el que se ubica en el primer lugar, para posteriormente tener dominancia el enfoque basado en las aristas y mantener una clasificación homogénea hasta el punto número 44, que es donde tiene lugar una medida de distancia entre

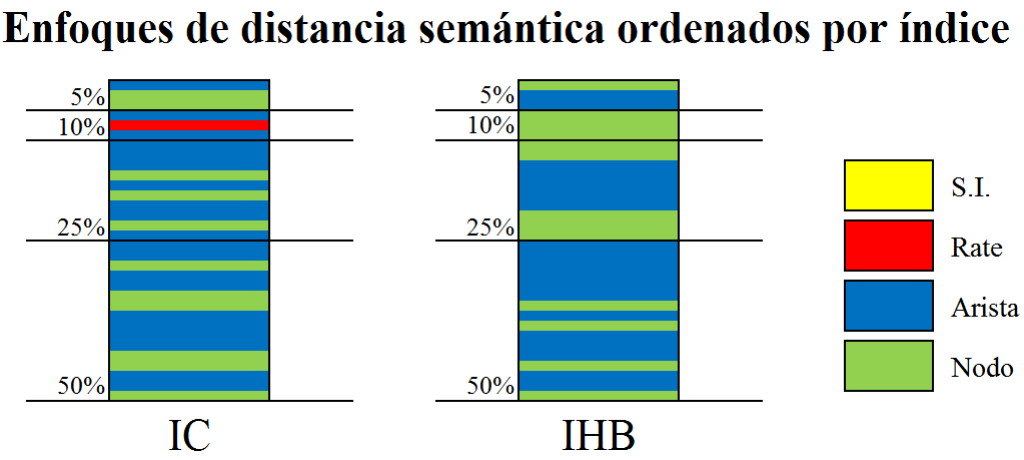

FIGURA 4.20: *Clasificación de los enfoques de similitud semántica (S.I., Nodos, Aristas y Tasa) clasificados por IC e IHB, para* $Experimento_{1-\rho}$. *La figura representa al* 50 % *mejor clasificado.*

genes en base a términos que no hace uso de relaciones semánticas entre ellos. Es particularmente interesante que la parametrización *S.I.*, es la peor evaluada cuando se analiza bajo el *IHB*. Los resultados son coherentes con los expuestos en la tabla 4.4, donde se observa que predomina la medida de $Dist_{tbk}$ predomina a pesar de que $Dist_{lin}$ tiene el mayor valor de *IHB*.

Para el experimento $Experimento_{1-|\rho|}$ se analiza la clasificación con respecto al *IC Absoluto* y al *IHB* (figura 4.21 para ver al 50 % mejor clasificado, y figura C.2 para ver la totalidad de la clasificación), obteniéndose un comportamiento similar al $Experimento_{1-\rho}$ pues, cuando se analiza la clasificación generada por el *IC Absoluto* se tiene una dominancia del enfoque basado en las aristas en los primeros lugares. Es interesante la buena evaluación de *Rate*, la cual se ve reflejada desde el tercer puesto clasificado. Para este experimento, la parametrización *S.I.* está bien evaluada respecto de la correlación de los perfiles de expresión, al presentarse en la posición número 24. Los resultados son coherentes con lo expuesto por la tabla 4.6, al indicar una clasificación similar entre $Dist_{tbk}$ y $Dist_{wp}$ (aristas) con $Dist_{lin}$ y $Dist_{jc_{norm}}$ (nodos), considerando que $Dist_{tbk}$ está presente en los primeras clasificaciones.

Para el caso de la clasificación con respecto al *IHB* (similar al $Experimento_{1-\rho}$) también se tiene una dominancia en los primeros puestos del enfoque basado en los nodos, para posteriormente dominar el enfoque basado en las aristas. El no uso de relaciones

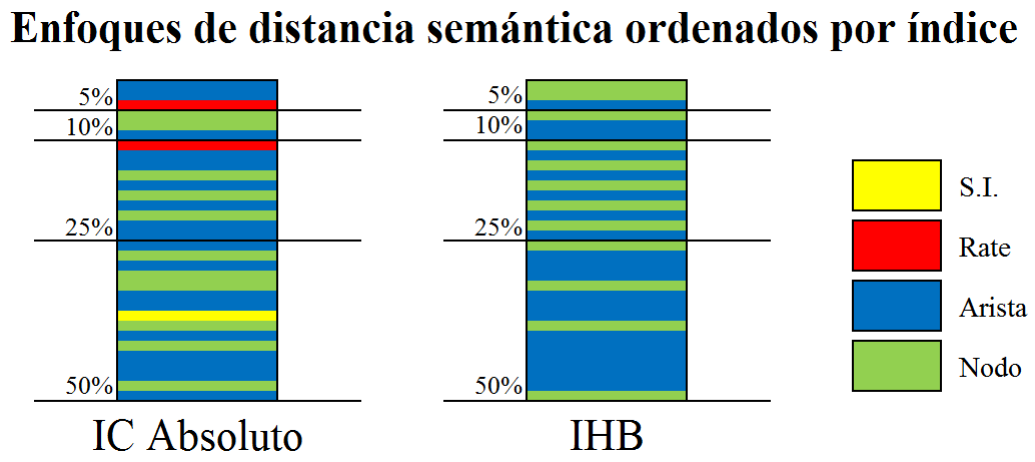

FIGURA 4.21: *Clasificación de los enfoques de similitud semántica (S.I., Nodos, Aristas y Tasa) clasificados por IC Absoluto e IHB, para* $Experimento_{1-|\rho|}$. *La figura representa al* 50 % *mejor clasificado.*

semánticas es mal evaluado, teniéndo su primera participación en la clasificación número 41. La parametrización *S.I.*, nuevamente hace su aparición en el último lugar de la clasificación. Los resultados tienen relación con lo expuesto en la tabla 4.7, pues si bien $Dist_{jc_{norm}}$ tiene la mejor evaluación con respecto a *IHB*, es la medida de $Dist_{tbk}$ (arista) la que se ve mayormente presente el conjunto de Pareto.

Hasta el momento los experimentos $Experimento_{1-\rho}$ y $Experimento_{1-|\rho|}$ no han presentado grandes diferencias, y se espera que el experimento $Experimento_{\rho+\rho}$ tuviera un comportamiento similar variando sólo en características aisladas, sin embargo, en la figura 4.22 se observa un comportamiento inverso (en la figura C.3 se presenta la totalidad de la clasificación), pues es el enfoque basado en los nodos el que entrega las mejores clasificaciones al medir la correlación de los comportamientos de los genes. La no incorporación de anotaciones biológicas es, además, particularmente bien evaluada, apareciendo en el puesto número nueve, lugar en que se comienza a manifestar una clasificación uniforme de los enfoques de similitud semántica. Estos resultados se materializan por lo expuesto en la tabla 4.9, donde las medidas de $Dist_{lin}$ y $Dist_{jc_{norm}}$ (nodos) se distribuyen uniformemente con respecto a $Dist_{lc}$ y $Dist_{tbk}$ (aristas), teniendo en consideración de que $Dist_{lin}$ presenta los valores mayores del *IC Absoluto*.

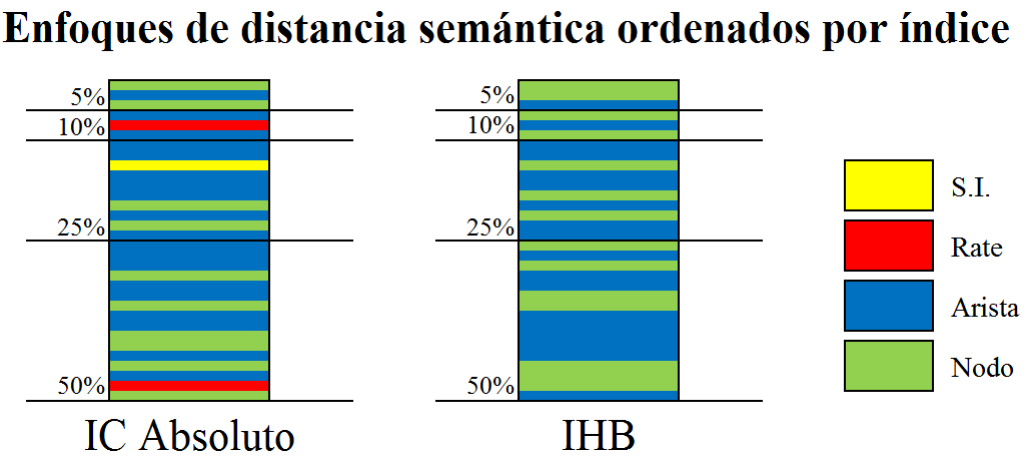

FIGURA 4.22: *Clasificación de los enfoques de similitud semántica (S.I., Nodos, Aristas y Tasa) clasificados por IC Absoluto e IHB, para* $Experimento_{\rho+\rho}$. *La figura representa al* 50 % *mejor clasificado.*

Para la clasificación respecto del *IHB* se mantiene el hecho de que el enfoque basado en los nodos entrega los mejores resultados, para posteriomente ser el enfoque basado en las aristas el que domina las clasificaciones (hasta la novena posición). Se mantiene el bajo efecto de una medida de distancia que no hace uso de similitudes semánticas (*Rate* tiene su primera clasificación en el puesto número 44) y la parametrización *S.I.* aparece en los últimos puestos (lugar número 61).

Con los análisis anteriores, se contesta la pregunta de análisis expuesta anteriormente:

4. ¿Cuál de los dos enfoques de medidas de distancia semántica se ve mayormente reflejado en las parametrizaciones mejor evaluadas, para las ontologías biológicas?

La respuesta es clara y varía por experimento y lo que se desee privilegiar, es decir, si se desea que los grupos generados tengan mayor correlación de sus perfiles de expresión o mayor homogeneidad en sus perfiles funcionales. En particular, la selección del enfoque debe responder a las reglas expuestas a continuación:

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $1-\rho$, utilizar un enfoque basado en las aristas para la similitud semántica entre términos biológicos si

se desea tener mayor correlación de los perfiles de expresión de los genes, o un enfoque basado en los nodos si se desea tener una mayor homogeneidad de los perfiles funcionales de los genes.

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $1-|\rho|$, utilizar un enfoque basado en las aristas para la similitud semántica entre términos biológicos si se desea tener mayor correlación de los perfiles de expresión de los genes, o un enfoque basado en los nodos si se desea tener una mayor homogeneidad de los perfiles funcionales de los genes.

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $\rho+\rho$, utilizar un enfoque basado en los nodos para la similitud semántica entre términos biológicos tanto si se desea tener mayor correlación de los perfiles de expresión de los genes como una mayor homogeneidad de los perfiles funcionales de los genes.

Los resultados permiten, además, confirmar lo expuesto en la sección 4.2.1 dada la coherencia de los resultados destacados por el criterio de Pareto que permitían reconocer a las mejores clasificaciones, puesto que los enfoques de dichos conjuntos de Pareto se ven también reflejados cuando se realiza una clasificación de las parametrizaciones, y se evalúa cuáles son los enfoques con mayor influencia al momento de maximizar los valores de los índices de validación.

4.2.5 Mejor medida de distancia semántica para anotaciones biológicas

Siguiendo la misma lógica de la sección 4.2.4, y para responder a la pregunta número cinco se lleva a cabo un análisis que permite identificar las medidas de distancias semánticas mejor clasificadas respecto de un índice de validación adecuado para los tres experimentos realizados ($Experimento_{1-\rho}$, $Experimento_{1-|\rho|}$ y $Experimento_{\rho+\rho}$). Nuevamente se clasifican las 63 parametrizaciones posibles por experimento del mayor al menor valor, identificando las

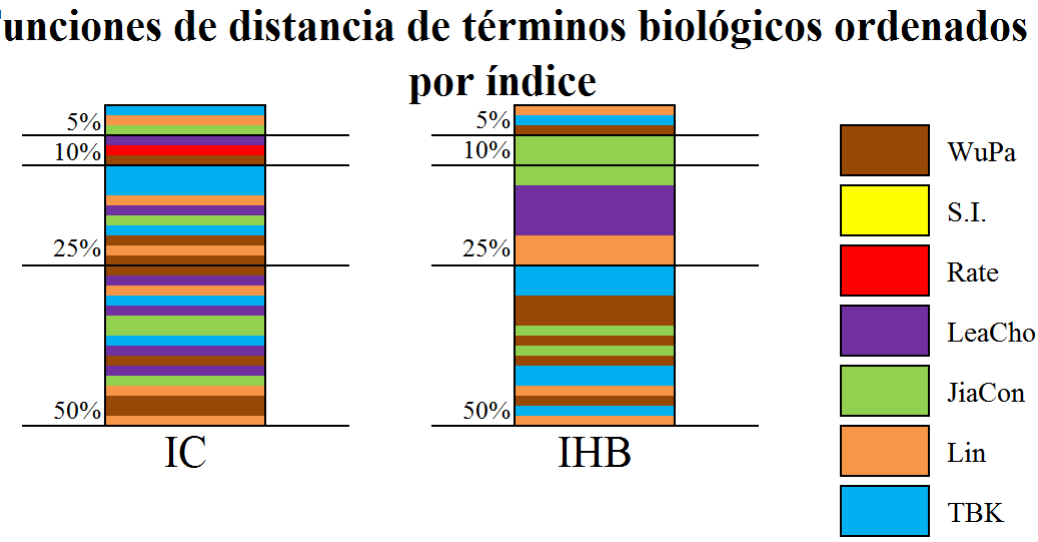

FIGURA 4.23: *Clasificación de medidas de distancia entre términos biológicos (S.I., Rate, $Dist_{wp}$, $Dist_{tbk}$, $Dist_{lc}$, $Dist_{jc_{norm}}$ y $Dist_{lin}$) clasificados por IC e IHB, para $Experimento_{1-\rho}$. La figura representa al 50 % mejor clasificado.*

medidas de distancia semántica entre términos biológicos que están presentes por cada valor de clasificación.

Para el caso del experimento $Experimento_{1-\rho}$, en la figura 4.23 (que presenta al 50 % mejor clasificado, mientras que la figura C.4 muestra la totalidad de la clasificación) se ve una dominancia compartida entre todas las medidas que establecen una distancia entre los términos biológicos (se incluye *Rate*, pues a pesar de no establecer una similitud semántica entre los términos, se encuentra presente dentro de la clasificación con respecto a los índices de validación), dejando en la posición número 38 a la parametrización *S.I.* Los resultados tienen relación con lo expuesto en la figura 4.20, pues si bien $Dist_{tbk}$ domina en los primeros lugares, $Dist_{lin}$ y $Dist_{jc_{norm}}$ en conjunto establecen la presencia del enfoque basado en los nodos.

Respecto del análisis de la clasificación realizada con el *IHB*, se observa la completa dominancia del enfoque basado en los nodos, pues $Dist_{lc}$, $Dist_{tbk}$ y $Dist_{wp}$ son menos representativas con respecto a $Dist_{jc_{norm}}$ y $Dist_{lin}$. Tal y como se a mencionado en análisis anteriores, la distancia que no considera similitud semántica se hace presente en último lugar de la clasificación cuando lo que se ha de medir, es la coherencia biológica de los grupos generados

por la parametrización.

Para el experimento $Experimento_{1-|\rho|}$, y clasificando del mismo modo que en el análisis anterior (pero esta vez utilizando el *IC Absoluto*), en la figura 4.24 (y figura C.5 para ver a la totalidad de la clasificación) se observa la dominancia de la distancia $Dist_{tbk}$, lo que confirma lo expuesto anteriormente por la figura 4.21, en que se aprecia que el enfoque basado en las aristas es la medida dominante, distribuyéndose el uso de *Rate* y del enfoque basado en los nodos (por medio de $Dist_{jc_{norm}}$ y $Dist_{lin}$) de manera uniforme. El uso de la parametrización *S.I.* se mantiene en la posición 24 de la clasificación, lo que da cuenta de que la distancia entre perfiles de expresión génica de $1 - |\rho|$, genera resultados positivos al ser evaluado por el *IC Absoluto*.

Para el caso del análisis a través del *IHB*, $Dist_{jc_{norm}}$ entrega los mejores resultados, para después ser junto a $Dist_{lin}$ las distancias dominantes de las mejores clasificaciones, lo que confirma lo expuesto por la figura 4.21 en que se ve la dominancia de los primeros lugares del enfoque basado en los nodos. En términos generales, al igual que en análisis anteriores, la medida de *Rate* no entrega buenos resultados al querer calcular la homogeneidad biológica de los grupos, y la parametrización *S.I.* es la peor evaluada.

Para el tercer y último análisis de las medidas de distancia entre términos dominantes de las mejores clasificaciones con respecto a un índice de validación, en el $Experimento_{\rho+\rho}$ (figuras 4.25 y C.6) se confirma el hecho de que cuando se desea maximizar el valor de la correlación entre los perfiles de expresión de los genes, son las medidas basadas en los nodos las mejor evaluadas ($Dist_{lin}$). *Rate* se ubica dentro de las mejor clasificadas (puesto número cinco) al igual que la parametrización *S.I.* (noveno lugar), comportamiento que si bien es diferente a lo esperado, responde principalmente a la dominancia de la medida de distancia entre perfiles de expresión de $\rho+\rho$, dado que al cambiar sólo esa variable, el comportamiento de la evaluación con respecto a *IC Absoluto* varía, por ejemplo, en relación a un experimento como $Experimento_{1-|\rho|}$.

Referente a la clasificación del $Experimento_{\rho+\rho}$ por medio del *IHB*, se confirma la

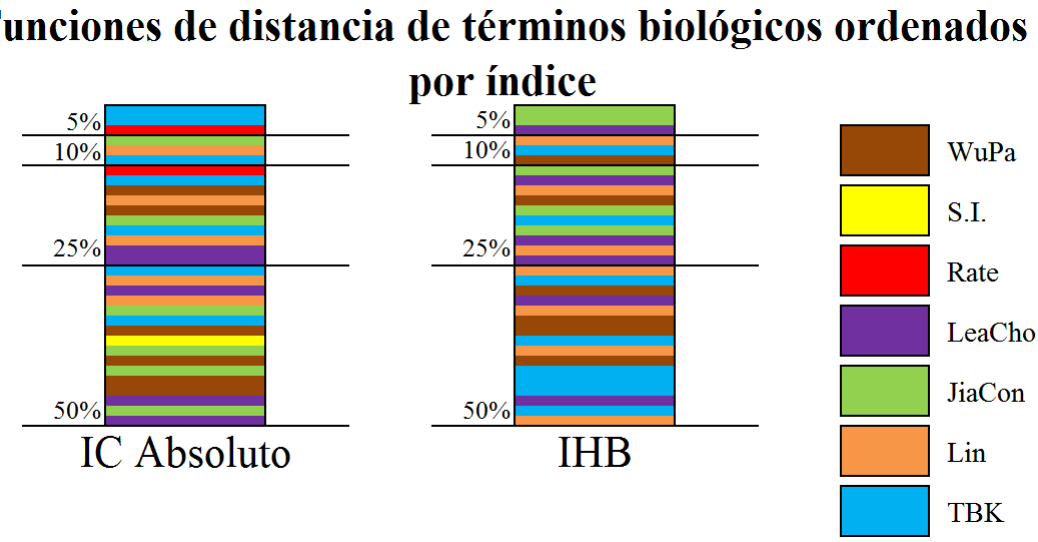

FIGURA 4.24: *Clasificación de medidas de distancia entre términos biológicos (S.I., Rate, $Dist_{wp}$, $Dist_{tbk}$, $Dist_{lc}$, $Dist_{jc_{norm}}$ y $Dist_{lin}$) clasificados por IC Absoluto e IHB, para $Experimento_{1-|\rho|}$. La figura representa al 50 % mejor clasificado.*

dominancia de las medidas de distancia de $Dist_{lin}$ y $Dist_{jc_{norm}}$, lo que justifica el hecho de que en la figura 4.22 indique que es efectivamente el enfoque basado en los nodos, el que entrega mejores resultados al querer incorporar conocimiento biológico a experimentos de expresión génica, considerando a la medida de correlación de $\rho + \rho$ como la utilizada para relacionar los perfiles de expresión de los genes en estudio.

Finalmente, y luego de la información entregada anteriormente, se entrega respuesta a la quinta pregunta de análisis, la cual dice:

5. ¿Cuál o cuáles medidas de similitud semántica se ven mayormente reflejada en las mejores parametrizaciones?

Nuevamente, y al igual que en la recomendación de un enfoque de distancia semántica entre términos biológicos, depende de lo que el experimentador desee que los genes consideren para relacionarse, y además de lo que espera que tenga mayor relevancia (si privilegiar a la información de expresión, o al conocimiento biológico). En términos generales, las mejores medidas de similitud semántica dependen del experimento, y son (para el conjunto de datos

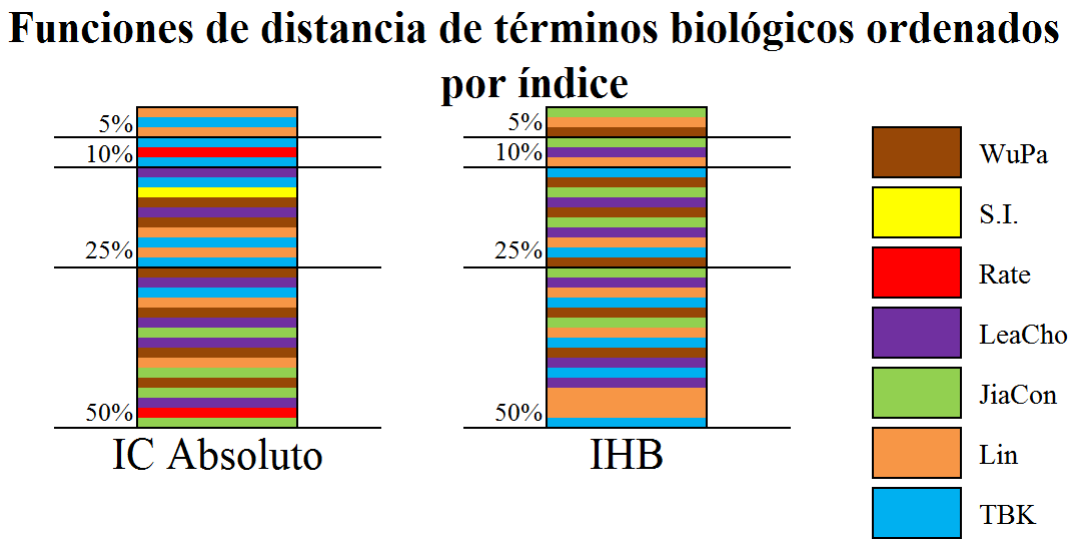

FIGURA 4.25: *Clasificación de medidas de distancia entre términos biológicos (S.I., Rate, $Dist_{wp}$, $Dist_{tbk}$, $Dist_{lc}$, $Dist_{jc_{norm}}$ y $Dist_{lin}$) clasificados por IC Absoluto e IHB, para $Experimento_{\rho+\rho}$. La figura representa al 50 % mejor clasificado.*

utilizado) las siguientes:

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $1-\rho$, utilizar la distancia de $Dist_{tbk}$ para la similitud semántica entre los términos biológicos si se desea tener mayor correlación de los perfiles de expresión de los genes, o las distancias de $Dist_{lin}$ y $Dist_{jc_{norm}}$ para la similitud semántica entre los términos biológicos si se desea tener una mayor homogeneidad de los perfiles funcionales de los genes, considerando que la distancia de $Dist_{lin}$ provee mejores resultados tanto para el *IC* como para el *IHB*, a diferencia de la distancia de $Dist_{jc_{norm}}$ que sólo provee buenos resultados para el *IHB*.

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $1-|\rho|$, utilizar la distancia de $Dist_{tbk}$ para la similitud semántica entre los términos biológicos si se desea tener mayor correlación de los perfiles de expresión de los genes, o la distancia de $Dist_{jc_{norm}}$ para la similitud semántica entre los términos biológicos si se desea tener una mayor homogeneidad de los perfiles funcionales de los genes.

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $\rho+\rho$, utilizar

la distancia de $Dist_{lin}$ para la similitud semántica entre los términos biológicos si se desea tener mayor correlación de los perfiles de expresión de los genes y además una mayor homogeneidad de los perfiles funcionales de los genes. Esta decisión se lleva a cabo dado que todas las medidas de similitud semántica entre términos comparten el primer lugar entre las mejores clasificadas, por lo que para discriminar cuál elegir, se considera la que además entrega el mejor resultado para el índice que mide la correlación de los perfiles de expresión de los genes de un grupo.

La presente sección también permite confirmar los resultados de tanto la sección 4.2.1, como la sección 4.2.4, dado que los resultados son coherentes tanto por la selección de un conjunto de buenas parametrizaciones maximizando funciones objetivo a través del criterio de Pareto, como con una clasificación utilizando los valores promedio de un índice de valiación.

4.2.6 Mejor función de distancia de perfiles funcionales

Con el objetivo de responder la sexta pregunta de análisis, se revisa para cada experimento ($Experimento_{1-\rho}$, $Experimento_{1-|\rho|}$ y $Experimento_{\rho+\rho}$) la clasificación de cada función de distancia entre los conjuntos de términos que describen a los genes para sus 63 parametrizaciones con respecto a un índice de validación adecuado, para así determinar cuál o cuáles son las funciones mejor evaluadas.

Al clasificar las funciones de distancia entre los perfiles funcionales de los genes para el experimento $Experimento_{1-\rho}$ (y por ende, a través del IC), tal y como se observa en la figuras 4.26 y C.7, la función PDP es notoriamente superior a las demás (lo cual se confirma al revisar los valores de la tabla 4.3). En términos generales, PDP, MDP y $Match$ entregan las mejores evaluaciones, lo que manifiesta que una selección de valores promedio con tendencia a valores mínimos de las distancias de los conjuntos de términos de los genes, proveen valores representativos de la distancia existente entre genes.

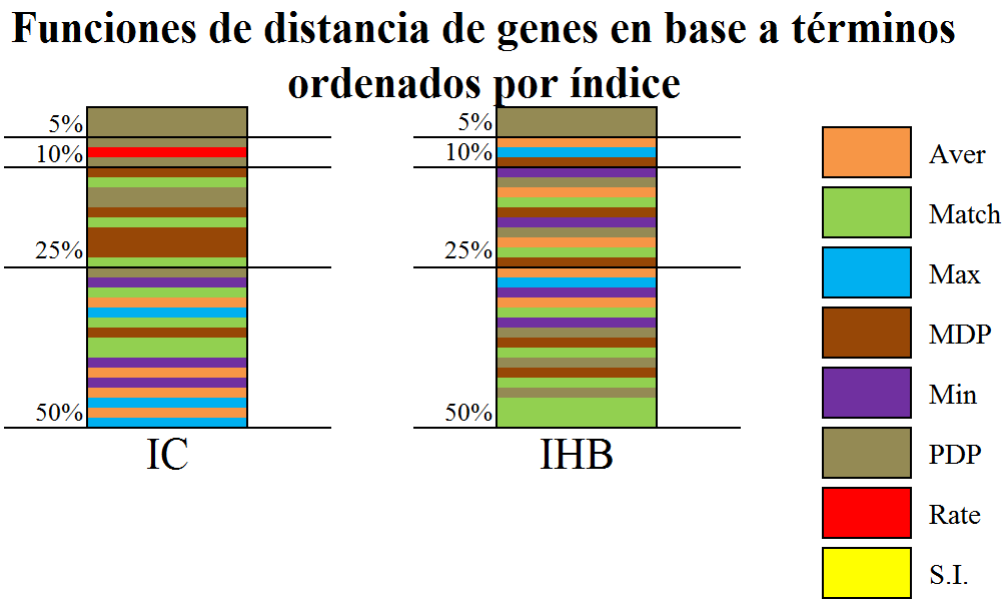

FIGURA 4.26: *Clasificación de medidas de distancia entre genes en base al conocimiento biológico (S.I., Rate, Min, Max, Aver, PDP, MDP y Match) clasificados por IC e IHB, para* $Experimento_{1-\rho}$. *La figura representa al* 50 % *mejor clasificado.*

Es interesante el hecho de que, cuando se clasifica a través del *IHB*, el comportamiento de la dominancia de las funciones es similar a cuando se clasifica con respecto al *IC*, donde además de las medidas *PDP*, *MDP* y *Match*, *Min* y *Aver* también presentan alta calificación. Lo anterior confirma el hecho de que, al menos para el $Experimento_{1-\rho}$, las medidas que no utilizan términos, o seleccionan valores representativos de los conjuntos con tendencia a los valores máximos, no entregan buenos resultados, siendo de hecho la medida *PDP* la mejor evaluada.

Al clasificar las parametrizaciones generadas con la medida de distancia entre perfiles de expresión de $1-|\rho|$ ($Experimento_{1-|\rho|}$) con respecto al *IC Absoluto*, se identifica nuevamente a *PDP* como la medida de distancia entre genes en base a sus perfiles funcionales mejor evaluada (observable en la figura 4.27 y corroborable con los valores expuestos en la tabla 4.8). Es importante notar, que al contrario de como lo indica la tabla 4.6, la función *Min* no entrega los valores más altos para *IC Absoluto*, por lo que se deduce que es seleccionada en el conjunto de Pareto más por su capacidad de separar los grupos, que por unir a los genes de acuerdo a sus

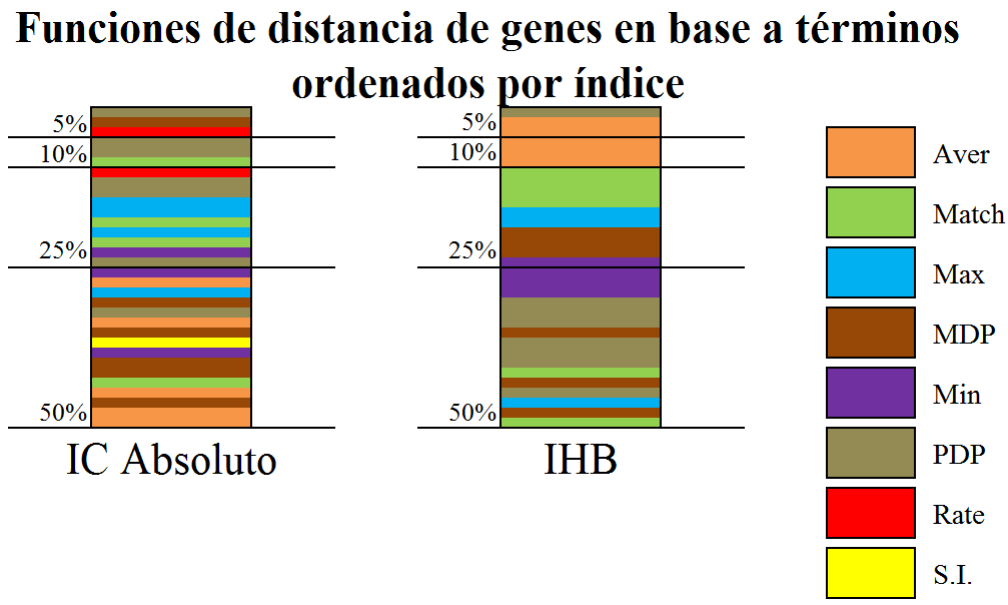

FIGURA 4.27: *Clasificación de medidas de distancia entre genes en base al conocimiento biológico (S.I., Rate, Min, Max, Aver, PDP, MDP y Match) clasificados por IC Absoluto e IHB, para* $Experimento_{1-|\rho|}$. *La figura representa al* 50 % *mejor clasificado.*

perfiles de expresión.

Cuando la clasificación se hace, en cambio, con respecto a los valores del *IHB*, a pesar de que *PDP* sigue siendo el mejor clasificado, se nota dominación de los primeros lugares de la medida que calcula el promedio entre todos los pares de distancias de los conjuntos de términos que describen a los genes (*Aver*). A pesar de ello, y como lo indica la tabla 4.7 es la medida *PDP* la que evidentemente se ve implicada en las parametrizaciones con mejores resultados, pues según lo visto en la sección 4.2.1 entrega una alta separación entre grupos.

Al igual que en análisis anteriores, la medida de distancia entre perfiles funcionales de $\rho+\rho$ presenta un comportamiento especial para los valores de las parametrizaciones que son mejor evaluados con respecto al *IC Absoluto* o al *IHB*. Contrario a lo esperado (pero coherente con lo expuesto en la tabla 4.9), las medidas mejor evaluadas son *PDP*, *Match* y *Max* (la cual, de hecho tiene un 50 % de participación del conjunto de Pareto del $Experimento_{\rho+\rho}$ evaluado con respecto al *IC Absoluto*), tal y como lo expone las figuras 4.28. A pesar de lo anterior, *PDP* sigue siendo dominante, incluso para el comportamiento diferente que manifiesta el $Experimento_{\rho+\rho}$

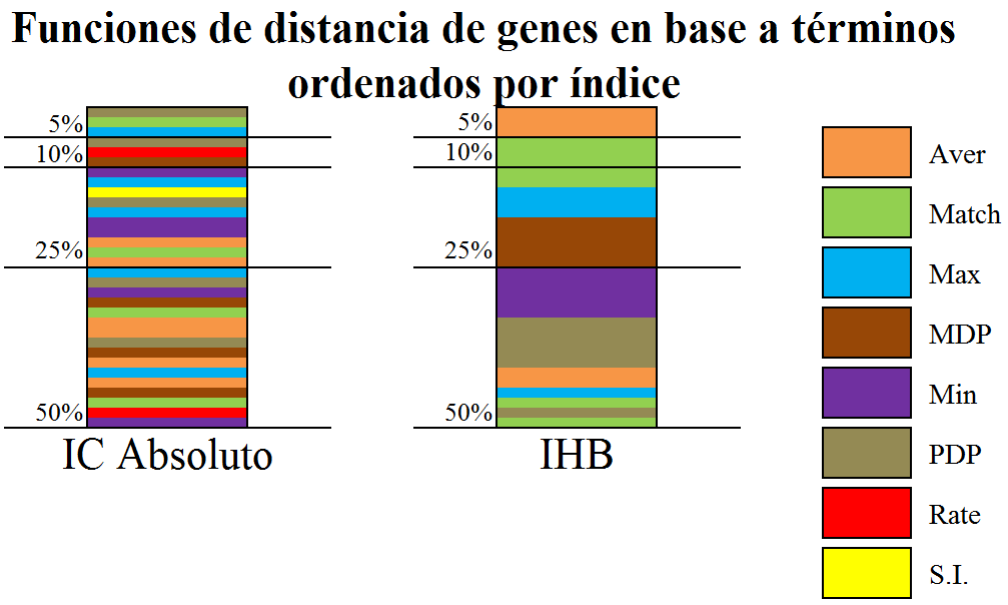

FIGURA 4.28: *Clasificación de medidas de distancia entre genes en base al conocimiento biológico (S.I., Rate, Min, Max, Aver, PDP, MDP y Match) clasificados por IC Absoluto e IHB, para Experimento$_{\rho+\rho}$. La figura representa al* 50 % *mejor clasificado.*

con el conjunto de datos.

Por último, al analizar los valores de *IHB* del *Experimento*$_{\rho+\rho}$, y clasificarlos de mayor a menor para las 63 parametrizaciones disponibles se tiene que todas las medidas se distribuyen uniformemente en los primeros lugares, dado que todos entregan el mismo valor de *IHB*. Analizar este caso resulta un tanto confuso, dado que el hecho de que los conceptos teóricos de *Max* y *Min* sean opuestos, no permite definir si los datos han de acomodarse a medidas que tienden a seleccionar valores mínimos, o valores máximos. Para llegar a una conclusión estable, se observan la tercera y cuarta medida predominante, las cuales al ser *Match* y *MDP*, denotan la tendencia de que seleccionar valores mínimos de distancias entre términos como representativos de la distancias entre dos genes en base a sus perfiles funcionales, presenta valores de evaluación altos al analizar la coherencia biológica de los grupos. A pesar de lo anterior, dado que *PDP* entrega buenos resultados al evaluar con respecto a *IC Absoluto*, su selección es teóricamente correcta cuando el experimento en cuestión utiliza la medida de correlación entre perfiles de expresión de $\rho + \rho$.

Como se ha hecho en los análisis anteriores, sólo resta contestar la pregunta de análisis relacionada a lo expuesto anteriormente, la cual dice:

6. ¿Cuál o cuáles parámetros asociados a la selección de una distancia representativa de los conjuntos de anotaciones biológicas de los genes en estudio se ven mayormente reflejados en las mejores parametrizaciones?

En este caso también ha de analizarse separadamente el comportamiento de las medidas de distancia entre genes en base a sus términos con respecto a los índices de validación de cada experimento realizado, pues éstos al buscar representar una relación diferente entre los genes, tienen también tendencia a funcionar mejor con ciertas medidas. En particular, las selecciones se resumen a continuación:

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $1-\rho$, utilizar la distancia *PDP* para la similitud entre los genes en base a sus perfiles funcionales tanto si se desea tener mayor correlación de los perfiles de expresión, como homogeneidad biológica de los perfiles funcionales de los genes.

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $1-|\rho|$, utilizar la distancia *PDP* para la similitud entre los genes en base a sus perfiles funcionales tanto si se desea tener mayor correlación de los perfiles de expresión, como homogeneidad biológica de los perfiles funcionales de los genes (también es válido utilizar la medida *Aver* para mejorar la calidad de los grupos de acuerdo a la homogeneidad biológica de los genes que los componen, teniendo en consideración que la calidad de la similitud en base a los perfiles de expresión no será adecuada).

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $\rho + \rho$, utilizar la distancia *PDP* para mayor similitud entre los genes en base a sus perfiles funcionales si se desea tener mayor correlación de los perfiles de expresión y además

mayor homogeneidad de los perfiles funcionales de los genes. Al igual que con la decisión de la medida de distancia semántica entre términos, esta se lleva a cabo considerando la medida de distancia entre perfiles funcionales de los genes que maximiza la correlación entre sus perfiles de expresión.

Esta sección, al igual que las anteriores, permite validar los resultados de la sección 4.2.1, y además confirmar el hecho de que establecer una relación entre genes en base a sus perfiles funcionales y perfiles de expresión provee resultados mayores tanto al evaluar la homogeneidad biológica de los grupos generados, como la correlación de la información relacionada al comportamiento de los genes.

4.2.7 Mejor función de combinación de matriz de perfil funcional y de expresión

Finalmente, y para responder a la séptima y última pregunta del análisis, se determina a las mejores funciones que permiten la incorporación del conocimiento biológico a los perfiles de expresión génica, a partir de la clasificación que se genera a través de los índices de validación implementados para las 63 parametrizaciones que caracterizan a cada experimento ($Experimento_{1-\rho}$, $Experimento_{1-|\rho|}$ y $Experimento_{\rho+\rho}$). De esta manera, y dependiendo lo que desee el investigador, se facilita la selección del parámetro de incorporación asegurando la obtención de resultados bien evaluados por los índices propuestos.

Al revisar los conjuntos de Pareto expuestos en la sección 4.2.1 del $Experimento_{1-\rho}$, se detectan ciertas tendencias de medidas mejor evaluadas, las cuales son corroboradas en el siguiente análisis. En la figura 4.29 (y C.10 para ver la totalidad de la clasificación), si bien se considera que la medida *Eucli* posee las mejores evaluaciones, la medida de $\alpha = 0,5$ es claramente superior en todo el conjunto de mejores evaluaciones (lo cual se corrobora además por los datos expuesto en la tabla 4.3.

Cuando por otro lado se clasifica con respecto al *IHB*, se ve una clara la dominancia

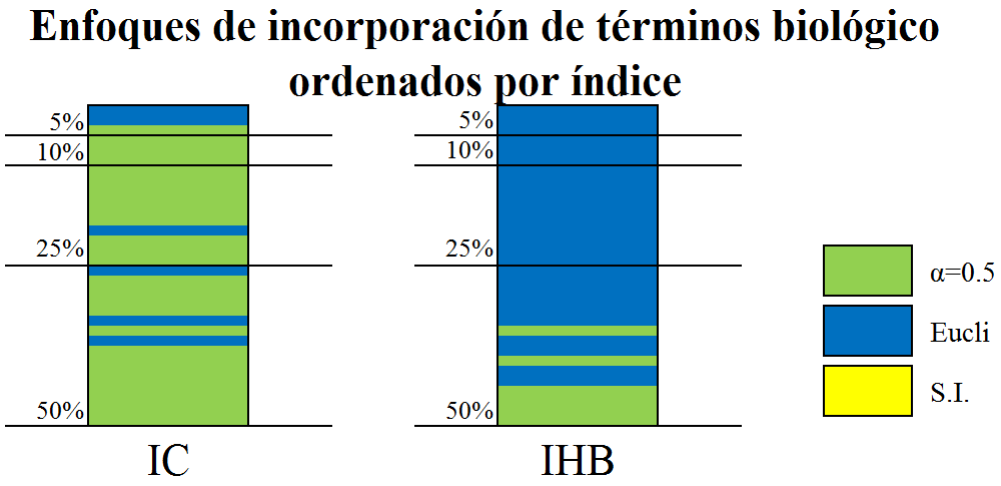

FIGURA 4.29: *Clasificación de funciones de incorporación de conocimiento biológico (S.I., $\alpha = 0,5$ y Eucli) clasificados por IC e IHB, para experimento $Experimento_{1-\rho}$. La figura representa al 50 % mejor clasificado.*

a lo largo de la mayoría de las mejores clasificaciones el uso de una medida *Eucli*, lo que es concordante con la información expuesta en la tabla 4.5, pero no con los datos de la tabla 4.4, donde el conjunto de parametrizaciones dominantes hacen uso de la medida $\alpha = 0,5$. De todos modos se considera que la medida *Eucli* es superior al resto cuando se requiere que los grupos posean mayor coherencia en sus perfiles funcionales.

Al analizar las parametrizaciones del $Experimento_{1-|\rho|}$ clasificadas según los valores del *IC Absoluto*, tal y como se observa en las figuras 4.30 y C.11, es dominante la medida de incorporación de $\alpha = 0,5$ en los primeros lugares mejor clasificados, lo cual es evidente al ver los valores de la tabla 4.6, donde todas las parametrizaciones del conjunto de Pareto utilizan dicha medida de distancia para incorporar el conocimiento biológico.

En lo que respecta a la clasificación a través del *IHB*, y similar a lo que ocurre en el experimento $Experimento_{1-|\rho|}$, se observa una dominancia de las mejores clasificaciones de una medida de combinación basada en *Eucli*, lo cual nuevamente no es del todo evidente en los datos de la tabla 4.7 pues sólo hay un $43,9\,\%$ de parametrizaciones que la utilizan. A pesar de ello, y dada su dominancia en el presente análisis, se considera que la medida *Eucli* es la mejor opción cuando se desea distanciar los genes en base a sus perfiles de expresión de acuerdo a la

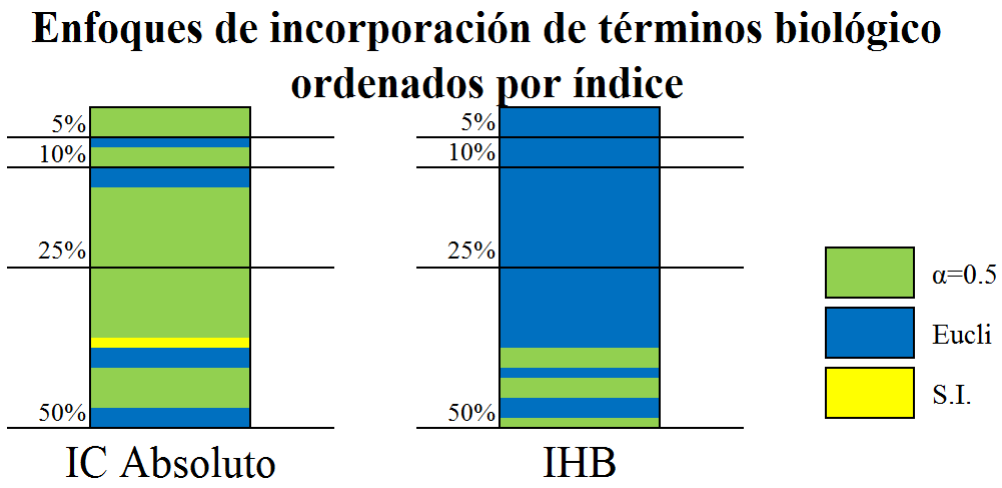

FIGURA 4.30: *Clasificación de funciones de incorporación de conocimiento biológico (S.I., $\alpha = 0,5$ y Eucli) clasificados por IC Absoluto e IHB, para experimento $Experimento_{1-|\rho|}$. La figura representa al 50 % mejor clasificado.*

medida $1 - |\rho|$, y se desea maximizar la coherencia biológica de los genes dentro de un grupo.

La medida de distancia entre perfiles de expresión de $\rho + \rho$ ha tenido un comportamiento diferente a lo esperado a lo largo de los análisis, sin embargo en lo que respecta a la inclusión de conocimiento biológico a los perfiles de expresión, hay concordancia con lo expuesto en los dos experimentos anteriores. De acuerdo a lo que indican las figuras 4.31 y C.12, se aprecia que para maximizar los valores de *IC Absoluto* se ha de privilegiar la elección de la medida $\alpha = 0,5$, lo cual además es coherente con lo expuesto en la tabla 4.9, donde todas las parametrizaciones que incorporan información utilizan dicha medida.

Para la clasificación de las parametrizaciones realizadas en el experimento $Experimento_{\rho+\rho}$, con respecto a los valores obtenidos para el *IHB*, se tiene que el uso de una medida *Eucli*, maximiza los valores de coherencia biológica en los grupos generados hecho que, nuevamente, no es del todo evidente en los datos de la tabla 4.10. Se debe tener presente el análisis realizado en la sección 4.2.1 referente a que, a pesar de que una medida *Eucli* entrega buenos resultados de *IHB* para el $Experimento_{\rho+\rho}$, esto se debe a que la combinación de ambos parámetros generan un único grupo con un alto valor del índice, por lo que no es teóricamente correcto seleccionar ambos parámetros si lo que se busca es justamente agrupar a la totalidad

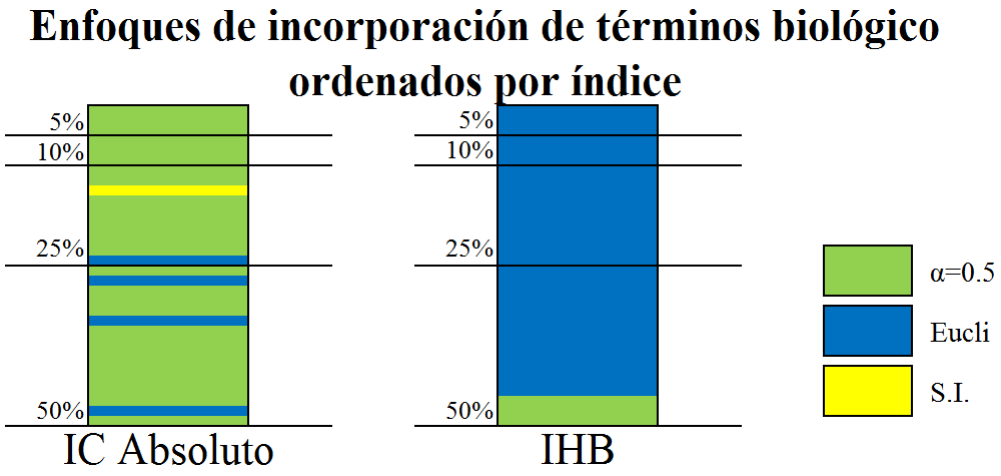

FIGURA 4.31: *Clasificación de funciones de incorporación de conocimiento biológico (S.I., $\alpha = 0,5$ y Eucli) clasificados por IC Absoluto e IHB, para experimento Experimento$_{\rho+\rho}$. La figura representa al 50 % mejor clasificado.*

de los genes en grupos de menor densidad. Dado lo anterior, a pesar del resultado numérico, se privilegia el uso de la medida $\alpha = 0,5$, con el argumento de que combinar la medida de correlación entre perfiles de expresión génica de $\rho + \rho$ con la medida de incorporación de conocimiento biológico *Eucli*, no genera una partición en el conjunto de datos.

Finalmente, para cerrar el ciclo de preguntas relacionadas al análisis de los resultados, se ha de contestar la séptima y última pregunta, la cual dice:

7. ¿Cuál de las dos funciones para la unión de matrices utilizada de ve reflejada con mayor frecuencia en las mejores parametrizaciones?

Para este caso, y dado que los tres experimentos se han alineado en comportamiento, se hace una conclusión de acuerdo a lo que el experimentador desee maximizar, es decir, si desea privilegiar la generación de grupos con mayor correlación de sus perfiles de expresión, o bien privilegiar la generación de grupos con mayor coherencia biológica de los perfiles funcionales de los genes que los conforman. Las recomendaciones de elección para una medida que permite la combinar ambos tipos de datos, son:

- Tanto si se desea que la coexpresión entre perfiles de expresión se realice mediante $1 - \rho$,

$1-|\rho|$, o $\rho+\rho$, utilizar la función $\alpha = 0,5$ para la incorporación del conocimiento biológico a los perfiles de expresión, si se desea tener mayor correlación de los perfiles de expresión en los grupos generados.

- Tanto si se desea que la coexpresión entre perfiles de expresión se realice mediante $1-\rho$ o $1-|\rho|$, utilizar la función *Eucli* para la incorporación del conocimiento biológico a los perfiles de expresión, si se desea tener mayor homogeneidad de los perfiles funcionales de los genes en los grupos generados.

- Si se desea que la coexpresión entre perfiles de expresión se realice mediante $\rho+\rho$, utilizar la función $\alpha = 0,5$ para la incorporación del conocimiento biológico a los perfiles de expresión, si se desea tener mayor homogeneidad de los perfiles funcionales de los genes en los grupos generados.

Para cerrar, el presente capítulo valida la totalidad del desarrollo al probar experimentalmente que la solución propuesta para la incorporación de anotaciones biológicas al algoritmo de agrupamiento *MST-kNN* (encontrando así relaciones entre genes), entrega resultados de mayor calidad que generando un agrupamiento que sólo considere perfiles de expresión génica. Resulta interesante que, además de obtener resultados beneficiosos desde el punto de vista de la coherencia biológica, la incorporación de anotaciones biológicas permite que la relación entre los genes mejore también la calidad de la correlación de sus perfiles de expresión (con parametrizaciones que entregan mejoras porcentuales de hasta un 242 %).

4.2.8 Resumen de los resultados

A partir de las recomendaciones de parametrización que se extraen de cada una de las respuestas a las siete preguntas expuestas en la sección 4.1.1, se establece el conjunto de las cinco mejores parametrizaciones clasificadas según el tipo de experimento a realizar y la variable

que se desee maximizar, pudiendo ser esta, la similitud de los genes de un grupo con respecto a sus perfiles de expresión génica, o la similitud de los genes de un grupo con respecto a sus perfiles funcionales. El detalle de las cinco parametrizaciones mencionadas, y que representan al respaldo teórico resultante de la presente tesis, es el siguiente:

- Si se desea establecer una medida de correlación entre los perfiles de expresión génica de los genes en base a $1 - \rho$, se tienen dos opciones de maximización de función objetivo:
  - Para maximizar el valor de la coexpresión entre los perfiles de expresión de los genes en los grupos generados, la parametrización recomendada que entrega los mejores resultados corresponde a:
    - Utilizar la distancia basada en la ecuación 3.5 (SLIMANI et al., 2008), para establecer la distancia entre términos biológicos en base a su similitud semántica.
    - Utilizar el promedio de las distancias ponderadas entre los términos de los conjuntos de los genes (ecuación 3.9), para establecer la distancia entre genes en base a sus perfiles funcionales.
    - Utilizar la función de combinación de las matrices basada en un ponderador α (ecuación 3.11) con un valor igual a $0,5$, para la incorporación del conocimiento biológico a los perfiles de expresión génica.
  - Para maximizar el valor de la coherencia biológica, o similitud de los perfiles funcionales de los genes en los grupos generados, la parametrización recomendada que entrega los mejores resultados corresponde a:
    - Utilizar la distancia basada en la ecuación 3.8 (LIN, 1998), para establecer la distancia entre términos biológicos en base a su similitud semántica.
    - Utilizar el promedio de las distancias ponderadas entre los términos de los conjuntos de los genes (ecuación 3.9), para establecer la distancia entre genes en base a sus perfiles funcionales.

- Utilizar la función de combinación de las matrices basada en una distancia euclídea (ecuación 3.12), para la incorporación del conocimiento biológico a los perfiles de expresión génica.

- Si se desea establecer una medida de correlación entre los perfiles de expresión génica de los genes en base a $1 - |\rho|$, se tienen dos opciones de maximización de función objetivo:

  - Para maximizar el valor de la coexpresión entre los perfiles de expresión de los genes en grupos generados, la parametrización recomendada que entrega los mejores resultados corresponde a:

    - Utilizar la distancia la distancia basada en la ecuación 3.5 (SLIMANI et al., 2008), para establecer la distancia entre términos biológicos en base a su similitud semántica.
    - Utilizar el promedio de las distancias ponderadas entre los términos de los conjuntos de los genes (ecuación 3.9), para establecer la distancia entre genes en base a sus perfiles funcionales.
    - Utilizar la función de combinación de las matrices basada en un ponderador α (ecuación 3.11) con un valor igual a $0,5$, para la incorporación del conocimiento biológico a los perfiles de expresión génica.

  - Para maximizar el valor de la coherencia biológica, o similitud de los perfiles funcionales de los genes en los grupos generados, la parametrización recomendada que entrega los mejores resultados corresponde a:

    - Utilizar la distancia basada en la ecuación 3.7 (JIANG & CONRATH, 1997), para establecer la distancia entre términos biológicos en base a su similitud semántica.
    - Utilizar el promedio de las distancias ponderadas entre los términos de los conjuntos de los genes (ecuación 3.9), para establecer la distancia entre genes

en base a sus perfiles funcionales.

- Utilizar la función de combinación de las matrices basada en una distancia euclídea (ecuación 3.12), para la incorporación del conocimiento biológico a los perfiles de expresión génica.

- Si se desea establecer una medida de correlación entre los perfiles de expresión génica de los genes en base a $\rho + \rho$, se tiene una única opción de parametrización que permite maximizar las dos funciones objetivo:

  - Para maximizar tanto el valor de la coexpresión entre los perfiles de expresión, como el valor de la coherencia biológica o similitud de los perfiles funcionales de los genes en los grupos generados, la parametrización recomendada que entrega los mejores resultados corresponde a:

    - Utilizar la distancia basada en la ecuación 3.8 (LIN, 1998), para establecer la distancia entre términos biológicos en base a su similitud semántica.
    - Utilizar el promedio de las distancias ponderadas entre los términos de los conjuntos de los genes (ecuación 3.9), para establecer la distancia entre genes en base a sus perfiles funcionales.
    - Utilizar la función de combinación de las matrices basada en un ponderador α (ecuación 3.11) con un valor igual a $0,5$, para la incorporación del conocimiento biológico a los perfiles de expresión génica.

4.2.9 Comparación de los resultados

El algoritmo *MST-kNN* original (que no incorpora anotaciones biológicas de GO) posee un grado determinado de similitud de expresión génica y de homogeneidad biológica en los grupos que genera, los cuales sirven como indicadores de calidad de referencia para definir si el desarrollo

propuesto (con sus 186 parametrizaciones posibles, y al cual designaremos "*MST-kNN-Bio*") entrega mejores o peores resultados. Para comparar *MST-kNN-Bio*, se usa el algoritmo *InteGO* propuesto por Marie Verbanck (VERBANCK et al., 2013), el cual corresponde a una propuesta relacionada al estado del arte del problema atacado en esta tesis, y que por su modelamiento sirve de referencia para comparar *MST-kNN-Bio*.

El *Software InteGO*, es una herramienta desarrollada en *R* (ROSS & ROBERT, 1996), que implementa un algoritmo no supervisado para integrar el conocimiento biológico de las anotaciones de GO, a datos de expresión génica, de manera que dos genes son cercanos si poseen perfiles de expresión génica similares, y al mismo tiempo perfiles funcionales similares. *InteGO* utiliza como algoritmo de agrupamiento a *K-medias* para generar grupos de genes de acuerdo a su distancia, los cuales son evaluados mediante dos índices de validación, uno que mide la co-expresión, y otro que mide la coherencia biológica de los genes pertenecientes a un grupo.

Los parámetros de entrada que requiere el algoritmo *InteGO* son: una matriz de genes con sus perfiles de expresión génica asociados, una matriz de genes con sus perfiles funcionales asociados, la cantidad de grupos que se desea como resultado, entre otros. El algoritmo utiliza el coeficiente de correlación de *Pearson* para establecer la distancia entre genes en base a perfiles de expresión, previo modelamiento en una matriz donde cada fila representa a un gen y cada columna una muestra de un experimento, de manera que el valor (i, j) representa al valor de expresión de la muestra j del gen i. Para el caso del modelamiento de los perfiles funcionales, se genera una matriz binaria donde cada fila representa a un gen y cada columna un término biológico de GO, de manera que si el gen i está anotado en el término j el valor (i, j) es igual a uno, y en el caso contrario cero. El proceso consiste en yuxtaponer ambas matrices de acuerdo al criterio de considerar una función biológica co-expresada, es decir, asociar un término a un par de genes sólo si estos están co-expresados según *Pearson*.

Una primera desventaja observable de *InteGO*, es la necesidad de explicitar la cantidad

de grupos que se han de generar, variable que como se ha descrito en la sección 4.2.2, no es influyente de la calidad de los grupos generados. Dado que no es justo comparar un agrupamiento de *InteGO* con *MST-kNN-Bio*, con un número diferente de grupos (pues se ven afectados los valores promedio de los índices de validación), se decide considerar el número de grupos resultante de *MST-kNN-Bio* como parámetro de *InteGO* de los grupos a generar. Esta desventaja tiene relación directa con el algoritmo *K-medias* utilizado por *InteGO*, por lo que se evalúa la cantidad de grupos con las cuales comparar ambas propuestas de la siguiente manera: Sí a través de *MST-kNN-Bio* se obtiene para una parametrización X_i, una cantidad de grupos Agr_i, ello implica que para *InteGO* se realiza la prueba para una cantidad de grupos de: $\{Agr_i - 1, Agr_i, Agr_i + 1\}$. Dado que *InteGO* utiliza la correlación de *Pearson* para relacionar genes según sus perfiles de expresión, el tipo de experimento al cual se asocian (de las 186 parametrizaciones) es al que utilizan la ecuación 3.1, y por tanto sólo tiene relación con dos parametrizaciones recomendadas, las cuales son:

- Parametrización recomendada N°1:
  - Utilizar la ecuación 3.1 $(1 - \rho)$ como medida de correlación entre los perfiles de expresión génica de los genes.
  - Utilizar la ecuación 3.8 (LIN, 1998), para establecer la distancia entre anotaciones biológicas en base a su similitud semántica.
  - Utilizar la ecuación 3.9 (*PDP*) para establecer la distancia entre genes en base a sus perfiles funcionales.
  - Utilizar la ecuación 3.12 (*Eucli*) para la incorporación del conocimiento biológico a los perfiles de expresión génica.
- Parametrización recomendada N°2:
  - Utilizar la ecuación 3.1 $(1 - \rho)$ como medida de correlación entre los perfiles de expresión génica de los genes.

- Utilizar la ecuación 3.5 (SLIMANI et al., 2008), para establecer la distancia entre anotaciones biológicas en base a su similitud semántica.
- Utilizar la ecuación 3.9 (*PDP*) para establecer la distancia entre genes en base a sus perfiles funcionales.
- Utilizar la ecuación 3.11 con un valor igual a $0,5$ ($\alpha = 0,5$), para la incorporación del conocimiento biológico a los perfiles de expresión génica.

Una desventaja no menor que fue detectada en *InteGO*, al intentar ejecutar el algoritmo con el conjunto de datos de la levadura *Saccharomyces cerevisiae* utilizados en las pruebas de *MST-kNN-Bio*, es que no puede generar un agrupamiento cuando existen genes que no tienen anotaciones biológicas asociadas. Diferente es el caso de *MST-kNN-Bio*, donde esa excepción es tolerada por la representación de distancias entre genes en base al conocimiento biológico. Este conflicto disminuye la utilidad de *InteGO*, pues se pierde la posibilidad de relacionar genes a los cuales se les desconoce información, con aquellos genes de los cuales se conoce información, o se tiene información parcial. Dado lo anterior, se tuvo que reducir la cantidad de genes a agrupar de 2.467, a sólo 1.200, con el objetivo de que *InteGO* pudiera generar grupos de genes correctamente. No hay que dejar de lado el hecho de que esta característica disminuye la posibilidad de ampliar la información conocida, hacia aquellos genes de los cuales hay poco conocimiento, punto que además implica que se debe ejecutar nuevamente *MST-kNN-Bio* con el nuevo conjunto de datos, para así obtener el número de grupos con los cuales ejecutar a *InteGO*, los que se presentan en la tabla 4.12.

Dados los resultados de la tabla 4.12, se establece la ejecución de seis instancias de *InteGO*, donde el parámetro de cantidad de grupos a generar posee los valores del conjunto: $\{41, 42, 43, 69, 70, 71\}$. Con lo anterior, la cantidad de grupos generados por las seis parametrizaciones se observa en la tabla 4.13.

Con los datos de las tablas 4.12 y 4.13 se establecen las primeras comparaciones, destacándose la diferencia entre el número de grupos generados (el cual es controlado para el

TABLA 4.12: Resultados de la ejecución de las dos parametrizaciones que se comparan con el algoritmo *InteGO*, ejecutados con un conjunto de datos de 1.200 genes de la levadura *Saccharomyces cerevisiae*. Se indica la cantidad de grupos generada tras el agrupamiento, la cantidad de grupos representativos para el cálculo del promedio del valor de los índices de validación y el porcentaje de ellos sobre el total de grupos generados.

Parametrización	núm. grupos	$\|Agr_{max}\|$	%
(*A*) $1-\rho$, Lin, *PDP* y *Eucli*	70	28	$40,0$
(*B*) $1-\rho$, T.B.K., *PDP* y $\alpha = 0,5$	42	23	$54,8$

TABLA 4.13: Resultados de la ejecución de *InteGO* con un conjunto de datos de 1.200 genes de la levadura *Saccharomyces cerevisiae*. Se entrega la cantidad de grupos generados, la cantidad de grupos representativos de la parametrización para el cálculo del promedio de los valores de los índices y el porcentaje de estos sobre el total de grupos generados.

Parametrización	núm. grupos	$\|Agr_{max}\|$	%
i1	41	10	$24,4$
i2	42	10	$23,8$
i3	43	10	$23,3$
i4	69	15	$21,7$
i5	70	15	$21,4$
i6	71	15	$21,1$

caso de *InteGO*) y el número de grupos que efectivamente es considerado para el promedio del valor de los índices de validación (conjunto Agr_{max} descrito en la sección 4.1.1, y representados por la tabla 4.1 y figura 4.2). Tal y como se observa, el porcentaje de grupos "útiles" para el criterio de evaluación utilizado es considerablemente menor en el caso de *InteGO*, al utilizar desde un $21,1\,\%$ como valor mínimo y un $24,4\,\%$ como valor máximo de los grupos generados, mientras que en *MST-kNN-Bio* los porcentajes son de un $40,0\,\%$ y un $54,8\,\%$. Lo anterior se debe a que *InteGO* genera una cantidad considerable de grupos con un único elemento (*singleton*, lo cual no ocurre al utilizar el algoritmo *MST-kNN*). En la tabla 4.14 se observa el porcentaje de *singleton* generado por *InteGO*, respecto de la totalidad de grupos existentes.

Tal y como lo expone la tabla 4.14, los porcentajes de grupos sin utilidad es considerable, generándose un $43,5$ como mínimo y un $51,2$ como máximo de *singleton*. Se les considera grupos "inútiles" bajo el criterio de que, al ser un algoritmo de agrupamiento la idea es justamente

TABLA 4.14: Porcentaje de grupos con un único elemento que se generan para las cinco instancias de prueba de *InteGO*, con el conjunto de 1.200 genes de la levadura *Saccharomyces cerevisiae*.

Parametrización	núm. *singleton*	%
i1	21	51, 2
i2	21	50, 0
i3	22	51, 2
i4	30	43, 5
i5	31	44, 3
i6	31	43, 7

agrupar genes, por tanto no tiene sentido un grupo que sólo tiene un gen. Además, no se puede medir la calidad del grupo a través de algún tipo de correlación entre los elementos para un *singleton*.

Una característica importante al observar el agrupamiento generado, tanto para el caso de *MST-kNN-Bio* como para las instancias de *InteGO*, es que se genera un grupo de gran cardinalidad (comparada al del resto de los grupos), los cuales se presentan en la tabla 4.15, junto con algunas de sus características.

TABLA 4.15: Cardinalidad, porcentaje sobre el total de genes agrupados y valores de *IC* e *IHB* para el grupo con mayor cardinalidad que se genera con las parametrizaciones recomendadas de *MST-kNN-Bio*, y las respectivas instancias de ejecución de *InteGO* con las cuales se contrastan.

Parametrización	Card. grupo mayor	%	IC	IHB
(*A*) $1-\rho$, Lin, *PDP* y *Eucli*	570	47, 5	0, 044	0, 744
i1	1.057	88, 1	0, 043	0, 789
i2	1.055	87, 9	0, 043	0, 789
i3	1.054	87, 8	0, 043	0, 789
(*B*) $1-\rho$, T.B.K., *PDP* y $\alpha = 0, 5$	273	22, 8	0, 045	0, 740
i4	883	73, 6	0, 040	0, 773
i5	883	73, 6	0, 040	0, 773
i6	883	73, 6	0, 040	0, 773

La tabla 4.15 indica que si bien en *MST-kNN-Bio* se genera un grupo de alta cardinalidad, el más grande no tiene más del 48 % del total de genes sometidos al algoritmo, mientras que para *InteGO* el grupo de mayor cardinalidad poseen de aproximadamente un 88 % y un 73, 6 %

del total de genes agrupados, lo que sumado al alto porcentaje de *singleton* indica que el agrupamiento, o el modelamiento de datos no es totalmente adecuado. La tabla 4.15 muestra que el índice que mide la correlación de los perfiles de expresión génica, tiene valores de: $0,040$, $0,043$, $0,044$ y $0,045$. Para el caso de la homogeneidad biológica de los genes del grupo, los valores son de: $0,740$, $0,744$, $0,773$ y $0,789$, lo cual responde al razonamiento presentado en la sección 4.2.2 donde se ve que grupos de alta cardinalidad presentan una coherencia biológica alta, de hecho, si se considera un único grupo con los 1.200 genes, el valor de *IHB* es de $0,795$ (el valor de *IC* para un grupo con los 1.200 genes es de $0,044$).

Finalmente, para comparar la calidad del agrupamiento generado por el desarrollo propuesto con el que se obtiene utilizando *InteGO*, se analiza el promedio de los valores de los índices de validación aplicados los grupos representativos que genera *InteGO*, sin contar el grupo de mayor cardinalidad. Luego, para que la comparación tenga sentido, se considera una cantidad igual de grupos generados por las parametrizaciones de *MST-kNN-Bio*, manteniendo una cantidad similar de genes. Lo anterior no es trivial, dado que *InteGO* distribuye a los genes en los grupos de manera poco uniforme, dejando una gran cantidad de grupos con un único gen (tabla 4.14), y una gran cantidad de genes en un único grupo (tabla 4.15). La tabla 4.16 muestra la comparación resultante, especificando el valor del *IC* y la mejora porcentual de las instancias de *InteGO* con respecto a las parametrizaciones de *MST-kNN-Bio* con que se comparó.

Al observar los valores de la tabla 4.16, se destaca que ambas parametrizaciones propuestas son, en promedio, superiores a su respectiva instancia de *InteGO* con que han de compararse. La parametrización (A) posee una mejora porcentual de un 49 % en los tres casos con que se compara, mientras que para la parametrización (B) la mejora es despreciable, teniendo incluso un valor por debajo del promedio en la instancia *i6* de *InteGO*. Con lo anterior se concluye que, tanto *MST-kNN-Bio* como *InteGO* entregan resultados similares (o inferiores para el caso de *InteGO*) cuando se comparan los promedios de los valores del *IC* para los grupos generados,

TABLA 4.16: Comparación del *MST-kNN-Bio* con *InteGO*, para el conjunto de 1.200 genes de la levadura *Saccharomyces cerevisiae*. Se indica el valor del *IC* para $\overline{Agr}_{max}$, descartando aquellos que tienen un único elemento y al grupo con mayor cardinalidad. Además se indica la mejora porcentual de las instancias de *InteGO* con respecto a las parametrizaciones de *MST-kNN-Bio* que le corresponden.

Parametrización	Cant. Genes (grupos)	IC	%
(*A*) $1-\rho$, Lin, *PDP* y *Eucli*	96 (9)	0,516	0
i1	99 (9)	0,346	−49
i2	99 (9)	0,346	−49
i3	99 (9)	0,346	−49
(*B*) $1-\rho$, T.B.K., *PDP* y $\alpha = 0,5$	151 (14)	0,434	0
i4	172 (14)	0,413	−5
i5	172 (14)	0,428	−1
i6	165 (14)	0,441	2

considerando sólo aquellos que maximizan el valor del índice, y descartando aquellos grupos con un único elemento, además del grupo con mayor cardinalidad.

En la tabla 4.17 se presentan datos similares a la tabla 4.16, pero para el caso del *IHB*, es decir, se entregan los valores promedio de cada parametrización e instancia, y las mejoras porcentuales de las parametrizaciones de *MST-kNN-Bio* contra las instancias de *InteGO* con que deben compararse. En ella se observa que las parametrizaciones recomendadas son superiores a las instancias de *InteGO* en todos los casos comparables, pero que esa superioridad no es significativa, por lo que se considera que ambos desarrollos (*MST-kNN-Bio* e *InteGO*) entregan resultados positivos cuando se analiza la homogeneidad biológia de los grupos generados.

Tanto *MST-kNN-Bio* como el *software InteGO*, entregan una solución al problema de integración de conocimiento biológico a un algoritmo de agrupamiento de genes que utiliza perfiles de expresión génica, a través del modelamiento de anotaciones biológicas de la base de datos GO y la combinación de esa información modelada con una representación de perfiles de expresión génica. Ambos desarrollos generan un agrupamiento de diferente calidad, donde se destaca que a *InteGO* se le debe hacer explícita la cantidad de grupos que se desea obtener, variable que como se describe en la sección 4.2.2 no respalda teóricamente la calidad del

TABLA 4.17: Comparación del *MST-kNN-Bio* con *InteGO*, para el conjunto de 1.200 genes de la levadura *Saccharomyces cerevisiae*. Se indica el valor del *IHB* para $\overline{Agr}_{max}$, descartando aquellos que tienen un único elemento y al grupo con mayor cardinalidad. Además se indica la mejora porcentual de las instancias de *InteGO* con respecto a las parametrizaciones de *MST-kNN-Bio* que le corresponden.

Parametrización	Cant. Genes (grupos)	IHB	%
(*A*) $1-\rho$, Lin, *PDP* y *Eucli*	91 (9)	0,901	0
i1	99 (9)	0,795	−13
i2	96 (9)	0,795	−13
i3	99 (9)	0,795	−13
(*B*) $1-\rho$, T.B.K., *PDP* y $\alpha = 0,5$	170 (14)	0,909	0
i4	172 (14)	0,832	−9
i5	172 (14)	0,831	−9
i6	165 (14)	0,847	−7

agrupamiento. *InteGO* produce una proporción de 43,5 % como mínimo y 51,2 % como máximo, de *singleton*, y además un grupo con una cardinalidad de 73,6 % como mínimo y 88,1 % como máximo del total de genes agrupados, ambas características disminuyen la utilidad del algoritmo pues, teóricamente, el agrupamiento tiene como objetivo el análisis de genes que compartan características y/o funciones similares, por tanto, carece de sentido el análisis de grupos donde sólo hay un gen, o de un grupo con una cardinalidad tal que no simplifica o reduce la complejidad del análisis de los elementos que lo conforman. Respecto del modelamiento de datos de *InteGO*, se rescata el hecho de que sólo considera términos en común entre dos genes si es que estos se co-expresan, condición que puede provocar un filtro importante del conocimiento biológico utilizado, además *InteGO* no considera la excepción de cuando se tiene genes sin descripción o perfil funcional, características que eliminan la posibilidad de extender el conocimiento adquirido sobre determinados genes, hacia aquellos de los cuales se les desconoce información. Al comparar *MST-kNN-Bio*, omitiendo las características negativas de *InteGO* descritas anteriormente, se tiene que al analizar los valores promedio del *IC* hay mejoras porcentuales sobre *InteGO* de un 1 % como mínimo y un 49 % como máximo, mientras que para el *IHB* la diferencia es menos significativa, alcanzándose mejoras porcentuales sobre *InteGO* de

un 7 % como mínimo y 13 % como máximo, lo cual da cuenta de que ambas propuestas cumplen el objetivo de generar grupos de genes cuyos perfiles de expresión génica estén correlacionados, y que además los grupos sean biológicamente coherentes.

Como resumen del presente capítulo, se estableció un proceso de cinco etapas para llevar a cabo la incorporación de anotaciones biológicas al algoritmo *MST-kNN* y validar los resultados, donde por cada uno de ellos se tenía una gama de funciones a implementar. Dado que la cantidad de parametrizaciones era de 186, fueron evaluadas en tres experimentos para así concluir cuáles parametrizaciones son las recomendadas, dependiendo de lo que el investigador tenga como objetivo. Cabe destacar que, a pesar de que cada análisis era independiente de los demás, todos entregaban resultados coherentes entre sí, lo que permite confirmar que el conjunto de parametrizaciones consideradas como las más adecuadas efectivamente entregan resultados satisfactorios, de acuerdo a los objetivos del experimento que se desee llevar a cabo. Para validar los resultados, se comparó el desarrollo propuesto tanto con el algoritmo *MST-kNN* original (que no incorpora anotaciones biológicas) como con un desarrollo perteneciente al estado del arte del problema atacado, siendo superior a ambos tanto en calidad de los resultados, como en el modelamiento y proceso de agrupamiento en general.

CAPÍTULO 5. CONCLUSIONES

En el campo de la bioinformática, las necesidades de la comunidad biológica relacionadas al análisis de grandes volúmenes de datos generados por herramientas computacionales se ha convertido en un elemento de estudio fundamental, puesto que el problema ya no pasa por cómo obtener los resultados de un experimento determinado, sino que en analizar y generar conclusiones respaldadas de la experimentación. Dentro de la teoría relacionada a la informática, la abstracción y el modelamiento de elementos propios de la biología permiten el uso de estructuras de datos y algoritmos que, aplicados a conjuntos de datos de un experimento, proveen una ayuda tanto del manejo de los datos, como del análisis de los mismos.

En particular, el uso de algoritmos de agrupamiento basados en grafos, han demostrado ser una herramienta potente cuando se trata de establecer relaciones entre conjuntos de datos de gran volumen. El algoritmo *MST-kNN*, es un algoritmo de agrupamiento aplicable a conjuntos de datos que requieran establecer relaciones entre genes en base a sus perfiles de expresión génica (INOSTROZA-PONTA et al., 2007). Si bien el algoritmo y su utilización demuestran ser un avance en el campo, se deja fuera una fuente de información biológica importante relacionada no sólo al comportamiento de los genes (como los perfiles de expresión), sino que a las funcionalidades que éstos tienen asociados.

Las anotaciones biológicas o términos biológicos que describen las funciones y características de un gen, permiten establecer relaciones entre genes a partir de información validada y confirmada por más de un estudio, lo cual la hace ser una fuente confiable deseable de utilizar. Es por ello que un objetivo no menor, es el de incorporar dicha variable de información al algoritmo de agrupamiento *MST-kNN*, de manera que sea capaz de relacionar genes entre sí no sólo en base a cómo éstos se comportan, sino que además tomando en cuenta las funciones y

características que puedan tener en común. En particular, el objetivo de la tesis es el de generar grupos de genes utilizando el algoritmo de agrupamiento *MST-kNN* que hagan uso tanto de los perfiles de expresión de los genes, como del conocimiento biológico externo relacionado a ellos.

En términos generales, la incorporación se lleva a cabo a través de cuatro etapas (y se evalúan posteriormente los resultados obtenidos en una quinta etapa), las cuales se describen en detalle en el capítulo tres, donde a través de la creación y combinación de matrices de distancia entre genes, se genera una representación de la similitud entre genes en base tanto a sus perfiles de expresión génica, como de sus perfiles funcionales determinados por las anotaciones biológicas de *Gene Ontology*.

Como se describe en detalle en la sección 4.1.1, por cada una de las cinco etapas del proceso de desarrollo hay un conjunto de funciones o métodos que se aplican, para lo cual se definen experimentos que permiten reconocer a las mejores parametrizaciones (o conjunto de parametrizaciones) de acuerdo a lo que un investigador desee, o necesite maximizar (o minimizar). Con lo anterior, los resultados presentes en este trabajo son adaptables de acuerdo a las necesidades del investigador. En particular se realizaron tres experimentos haciendo uso del conjunto de datos de la levadura *Saccharomyces cerevisiae* (EISEN et al., 1998) para los perfiles de expresión y *Gene Ontology* (GO, 2013) para los perfiles funcionales de los genes. En cada experimento se reconoce a las parametrizaciones que maximizan una característica específica de los grupos que se generan, obteniéndose como resultado cinco parametrizaciones (detalladas en la sección 4.2.8) adaptables al tipo de experimento que se desee y a la variable que se desee maximizar, pudiendo ser esta la similitud de los genes de un grupo en base a sus perfiles de expresión génica, o la coherencia biológica de los genes de un grupo en base a sus perfiles funcionales.

Una conclusión interesante de exponer, es el hecho de que se cuestionaba la efectividad de una medida de similitud semántica de un enfoque basado en las aristas para entregar resultados satisfactorios al ser aplicada a datos de una ontología biológica, sin embargo esta

se ve presente como recomendación en algunas parametrizaciones, cuando se desea obtener resultados satisfactorios en la correlación de los perfiles de expresión de los genes. Lo anterior puede deberse a que para el conjunto de datos utilizado, los términos biológicos puede estar distribuidos de manera uniforme en el DAG de GO; porque la aseveración de que el enfoque basado en las aristas tendrá resultados no satisfactorio es errónea, o bien porque la medida de similitud semántica propuesta por (SLIMANI et al., 2008) no presenta deficiencias al trabajar con los datos de GO.

Se debe tener presente además la conclusión relacionada a que la cantidad de grupos generados por una parametrización, haciendo uso del algoritmo *MST-kNN*, no es una función objetivo que deba maximizarse o minimizarse, dado que no se tiene un comportamiento predecible de la calidad del agrupamiento a partir de la cantidad de grupos a generar. Esta conclusión afecta directamente a desarrollos que utilicen algoritmos de agrupamiento que impliquen definir el número de grupos que se desea obtener, pues no se tiene una base teórica fundamentada para la selección de dicho número, cuando se tiene como objetivo maximizar la correlación de los genes pertenecientes a un grupo.

En términos globales, y haciendo un recuento de los objetivos específicos del trabajo realizado (sección 1.4.2), se tiene que el primer objetivo, de "conocer el funcionamiento del algoritmo *MST-kNN* a nivel teórico y de implementación, para realizar cambios en el mismo" se le da completitud en la sección 2.3 (apéndice A.2); al segundo objetivo de "identificar las bases de datos públicas de anotaciones biológicas, precisar sus características y revisar en la literatura métodos para representar el contenido extraído de ellas" se le da completitud en las sección 2.2 (apéndice B); del tercer objetivo de "desarrollar un módulo de *software* que implemente las modificaciones al algoritmo *MST-kNN*, incorporando las anotaciones biológicas modeladas", se muestra su funcionamiento lógico y procedural en la sección 3 (apéndice B.1.3), y el cuarto y último objetivo específico de "validar resultados a través de la prueba del módulo de *software* desarrollado sobre el conjunto de datos de la levadura *Saccharomyces cerevisiae*

(EISEN et al., 1998) y las fuentes de anotaciones génicas almacenadas en *Gene Ontology* (GO, 2013)", corresponde a todo el contenido expuesto en el capítulo cuatro.

Cabe señalar que la incorporación del conocimiento biológico al algoritmo de agrupamiento *MST-kNN*, no está atado sólo a la base de datos GO, pues puede ser aplicada a cualquier base de datos que ordene las anotaciones biológicas como un DAG o como un árbol, ya que esa característica permite el uso de medidas de similitud semántica basadas ya sea en el conteo de la cantidad de aristas que separan a las anotaciones biológicas entre sí, como en el análisis del contenido de información de las anotaciones biológicas en evaluación.

Como conclusión global, y en base a lo expuesto en la hipótesis de la tesis que dice: "dada la capacidad de agrupamiento del algoritmo *MST-kNN*, es posible añadir como parámetro del mismo, anotaciones biológicas de la levadura *Saccharomyces cerevisiae* disponibles en bases de datos de libre acceso, para encontrar grupos que estén asociados tanto a nivel de la expresión génica del organismo, como de sus anotaciones biológicas", se confirma su validez y por tanto es aceptada, y más aun, complementada con el hecho de que la incorporación de conocimiento biológico estático al algoritmo de agrupamiento *MST-kNN*, entrega resultados superiores al analizar la calidad de los grupos que se generan a nivel de correlación de los perfiles de expresión de los genes contenidos en ellos, obteniéndose una mejora porcentual máxima de 93 %. Respecto a la mejora a nivel de homogeneidad biológica de los perfiles funcionales de los genes contenidos en los grupos generados, tras la utilización de *MST-kNN* en la incorporación de anotaciones biológicas, se obtienen mejoras porcentuales (para las parametrizaciones recomendadas) de un 80 % como mínimo y un 117 % como máximo, siendo esta segunda variable altamente superior al algoritmo original y lo que da cuenta de que la incorporación de anotaciones biológicas al algoritmo de agrupamiento *MST-kNN* permite, efectivamente, la generación de grupos de genes que son buenos candidatos a análisis posteriores, punto que se ve reforzado al realizar una comparación del desarrollo propuesto con otro algoritmo relacionado al estado del arte del problema atacado (VERBANCK et al., 2013). El *software InteGO* también incorpora

anotaciones biológicas de GO a un algoritmo de agrupamiento no supervisado, pero al comparar su funcionamiento con el desarrollo propuesto en esta tesis, se evidencian las desventajas de tener que explicitar la cantidad de grupos que se desea como salida, de no poder trabajar con genes que no tengan asociado un perfil funcional o conjunto de anotaciones biológicas que los describan, de generar un $43,5\,\%$ como mínimo y un $51,2\,\%$ como máximo de grupos que sólo contienen un elemento y de generar un grupo con una cardinalidad de aproximadamente un $88\,\%$ y un $73,6\,\%$ del total de genes agrupados. La desventaja asociada a tener que indicar la cantidad de grupos que se requiere como salida es compleja, pues esa decisión no se puede tomar esperando optimizar la calidad de los resultados, pues como se indicó en la sección 4.2.2, la cantidad de grupos obtenidos no tiene relación con la calidad del agrupamiento, de todos modos se debe tener presente que esa desventaja está asociada al algoritmo utilizado por *InteGO* (*K-Medias*) y no al algoritmo mismo. La desventaja de *InteGO* de generar una gran cantidad de grupos con un elemento, y un grupo con un gran número de elementos se contradicen con el objetivo de generar grupos que faciliten el análisis de genes con características similares. Por otro lado, la desventaja de no soportar genes que no tengan anotaciones biológicas asociadas, no permiten la extensión del conocimiento desde elementos conocidos y validados, hacia los cuales se les desconoce información. Además de las diferencias anteriormente descritas, el desarrollo propuesto presenta mejoras porcentuales respecto de *InteGO* de un $1\,\%$ como mínimo y un $49\,\%$ como máximo cuando se evalúa la similitud de los perfiles de expresión génica, y de un $7\,\%$ como mínimo y un $9\,\%$ como máximo cuando se evalúa la coherencia biológica de los grupos generados, para un conjunto de datos adaptado a lo que *InteGO* requiere, y eliminando tanto los grupos con un único elemento, como el grupo de mayor cardinalidad (para poder hacer comparaciones equitativas).

La importancia de la solución implementada en la presente tesis radica en el uso de la información biológica para establecer relaciones entre genes, la cual al ser combinada con un tipo de dato ampliamente utilizado en el área (expresión génica), permiten establecer relaciones

entre genes con un respaldo científico mayor a que si sólo se considerara una de las fuentes de información mencionadas. La incorporación de información biológica en sí misma, representa un avance significativo para el establecimiento de relaciones o similitudes entre genes, pudiendo traducirse en la asociación de funciones para genes con una vaga descripción biológica, y que puede ser un puente para, por ejemplo, desarrollar drogas que combatan alguna enfermedad, o reconocer al gen asociado a una patología específica.

5.1 TRABAJO FUTURO

La presente tesis abre una serie de posibilidades de nuevos estudios que son interesantes de llevar a cabo. A continuación se enuncian algunos de los trabajos que pueden hacer uso del presente desarrollo, utilizándolo como un punto de inicio o base teórica para otros trabajos o experimentos interesantes de realizar.

El trabajo futuro puede relacionarse directamente con los alcances del trabajo (sección 1.4.3), así pues un primer desarrollo pendiente es la interpretación biológica de los resultados obtenidos, para con ello validar con un estudio acabado si las nuevas relaciones encontradas entre los genes de la levadura *Saccharomyces cerevisiae* responden a características coherentes con la especie, lo que permite revelar nuevos comportamientos o funcionalidades de la misma, o incluso respaldar estudios llevados a cabo por otros investigadores, fortaleciéndolos.

Un segundo trabajo interesante de realizar, es utilizar la solución propuesta sobre un conjunto de datos relacionados al ser humano (con su correspondiente interpretación biológica). Gran parte de los estudios biológicos o bioinformáticos tienen como motivación revelar características, funcionalidades, curas a enfermedades o comportamientos relacionados al ser humano, y es por ello que el desarrollo de la presente tesis, dada la calidad de los resultados obtenidos, abre la posibilidad de ser un aporte para el área o a investigaciones relacionadas a, por ejemplo, estudios sobre el cáncer.

Una característica no menor, es el hecho de haber incorporado las anotaciones biológicas a nivel de parámetro del algoritmo de agrupamiento *MST-kNN* (INOSTROZA-PONTA et al., 2007). Un trabajo futuro interesante, es utilizar la matriz de distancia entre genes en base a los perfiles funcionales y con ello generar una estructura (por ejemplo, de grafo) que sea incorporada a nivel interno del algoritmo. Así por ejemplo, tal y como el algoritmo elimina aristas de las estructuras *MST* y *kNN* generadas a partir de una matriz de distancia entre genes en base a perfiles funcionales, se puede llevar a cabo esa intersección, u otra de otro tipo, para las mismas iteraciones, u otras determinadas por la influencia que se quiera dar al conocimiento biológico, con la estructura de grafo o árbol generada a partir de la matriz de distancia entre genes en base a sus perfiles funcionales. Una variante de la idea anterior, es la de aplicar el algoritmo *MST-kNN* no directamente sobre la matriz que incorpora las anotaciones biológicas a los datos de expresión génica, sino que modificar el algoritmo para que reciba dos matrices de distancia, una basada en los perfiles de expresión de los genes ($\mathcal{M}_{expr}$) y otra basada en los perfiles funcionales de los genes ($\mathcal{M}_{func}$), de esa forma, la estructura *MST* puede ser generada a partir de $\mathcal{M}_{expr}$ mientras que la estructura *kNN* sea generada en base a la información contenida en $\mathcal{M}_{func}$ para luego continuar con el proceso de intersección de estructuras e iteraciones del algoritmo original. Claramente, se puede plantear la variante de que la estructura *MST* se genere a partir de los datos contenidos en $\mathcal{M}_{func}$, mientras que *kNN* sea generada a partir de la estructura $\mathcal{M}_{expr}$, para con ello generar los grupos de genes relacionados entre sí. Esa variante permite la incorporación del conocimiento biológico, pero desde otro enfoque que evite el uso del parámetro que combina ambas matrices de distancia (ponderador α o distancia euclídea).

REFERENCIAS

BENABDERRAHMANE, S., SMAIL-TABBONE, M., POCH, O., NAPOLI, A., & DEVIGNES, M.-D. (2010). **IntelliGO: a new vector-based semantic similarity measure including annotation origin**. En *BMC Bioinformatics*, tipo Methodology article. 54506 Vandoeuvre-lès-Nancy Cedex & 67404 Illkirch Strasbourg, Francia: BioMed Central Ltd. 11:588; DOI:10.1186/1471-2105-11-588; 1 de Diciembre.

CHERNOMORETZ, A. (2010). **Gene Ontology guided clustering of gene expression profiles**. En *Comprendio en Conferencia ISCB Latin America 2010*. Lunes 16 de Marzo. Montevideo, Uruguay.

CLARK, M. B., L.JOHNSTON, R., INOSTROZA-PONTA, M., FOX, A. H., FORTINI, E., MOSCATO, P., DINGER, M. E., & MATTICK, J. S. (2012). **Genome-wide analysis of long noncoding RNA stability**. En *Genome Research*, vol. 22, (pág. 885–898). Cold Spring Harbor Laboratory Press, 5 ed. DOI: 10.1101/gr.131037.111; Mayo.

COHEN, J. (2004). **Bioinformatics - an introduction for computer scientists**. En *ACM Computing Surveys*, vol. 36 de *No. 2*, (pág. 122–158). Junio.

DATTA, S., & DATTA, S. (2006). **Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes**. En *BMC Bioinformatics*, vol. 7 de *397*. University of Louisville, Louisville, KY 40202, USA: Department of Bioinformatics and Biostatistics. DOI: 10.1186/1471-2105-7-397; 31 de Agosto.

DREW, R. C. M., GANDHI, S. K., MCKAY, C. F., B., M. C., BERRETTA, R., YAHYA, V. S.,

INOSTROZA-PONTA, M., A., B. S., HEARD, N. R., VUCIC, S., STEWART, G. J., W., W. D., SCOTT, J. R., LECHNER-SCOTT, J., BOOTH, R. D., & MOSCATO, P. (2010). **A Transcription Factor Map as Revealed by a Genome-Wide Gene Expression Analysis of Whole-Blood mRNA Transcriptome in Multiple Sclerosis**. En *PLoS ONE*, vol. 5. E14176.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., & BOTSTEIN, D. (1998). **Cluster analysis and display of genome-wide expression patterns**. En *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, (pág. 14863–14868). Diciembre.

GESCHWIND, D. H., & KONOPKA, G. (2009). **Neuroscience in the era of functional genomics and systems biology**. En *Nature - International weekly journal of science*, vol. 461 de *No. 7266*, (pág. 908–915). 14 de Octubre.

GO (2013). **Gene Ontology**. Último acceso: Junio 2013.
URL `http://geneontology.org/`

GONZÁLEZ-BARRIOS, J. M., & QUIROZ, A. J. (2003). **A clustering procedure based on the comparison between the k nearest neighbors graph and the minimal spanning tree**. En *Statistics & Probability Letters, Elsevier*, vol. 62 de *1*, (p. 23–34). Copyright ©2003 Elsevier Science B.V. DOI: 10.1016/S0167-7152(02)00421-2; 15 de Marzo.

INOSTROZA-PONTA, M. (2008). **An Integrated and Scalable Approach Based on Combinatorial Optimization Techniques for the Analysis of Microarray Data**. School of Electrical Engineering and Computer Science. Thesis (Ph.D.), University of Newcastle. Australia.

INOSTROZA-PONTA, M., BERRETTA, R., & MOSCATO, P. (2011). **QAPgrid: A Two**

Level QAP-Based Approach for Large-Scale Data Analysis and Visualization. En *PLoS ONE*, vol. 6 de *No. 1*. DOI:10.1371/journal.pone.0014468; e14468.

INOSTROZA-PONTA, M., MENDES, A., BERRETTA, R., & MOSCATO, P. (2007). **An integrated QAP-based approach to visualize patterns of gene expression similarity**. En . J. W. M. RANDALL, H. A. ABBASS (Ed.) *Actas de la III Conferencia Australiana Progress in Artificial Life*, vol. 4828 de *Lecture Notes in Computer Science*, (pág. 156–167). ACAL, 978-3-540-76930-9, Diciembre 4-6, Gold Coast, Australia.

JAIN, A. K., MURTY, M. N., & FLYNN, P. J. (1999). **Data clustering: a review**. En *ACM Computing Surveys*, vol. 31 de *3*, (pág. 264–323). ACM New York, NY, USA. DOI: 10.1145/331499.331504; ISSN: 0360-0300; EISSN: 1557-7341; Septiembre.

JIANG, J. J., & CONRATH, D. W. (1997). **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy**. En *Proceedings of the International Conference on Research in Computational Linguistics ROCLING X*, (pág. 19–33). Taiwan. 20 de Septiembre.

LEACOCK, C., & CHODOROW, M. (1998). **Combining Local Context and WordNet Similarity for Word Sense Identification**. En *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, (pág. 265–283). MIT Press. Cambridge, Massachusetts. London, England: MIT Press. Christiane Fellbaum. ©1998 Massachusetts Institute of Technology, 1ra ed. ISBN: 0-262-06197-X; Mayo, 1998.

LIN, D. (1998). **An Information-Theoretic Definition of Similarity**. En *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning*, (pág. 296–304). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ©1998. ISBN: 1-55860-556-8.

LOCKHART, D. J., & WINZELER, E. A. (2000). **Genomics, gene expression and DNA**

arrays. En *Nature - International weekly journal of science*, vol. 405 de *No. 6788*, (pág. 827–836). 15 de Junio.

MARAZIOTIS, A. I., DIMITRAKOPOULOS, G., & BEZERIANOS, A. (2012). **Gene Ontology Semi-supervised Possibilistic Clustering of Gene Expression Data**. En *7th Hellenic Conference on AI, SETN 2012. Proceedings*, vol. 7297, (pág. 262–269). Lamia, Greece: Springer-Verlag Berlin Heidelberg, 1ra ed. DOI: 10.1007/978-3-642-30448-4_33; Print ISBN: 978-3-642-30447-7; Online ISBN: 978-3-642-30448-4; Series ISSN: 0302-9743; 28-31 de Mayo.

MEYER, F., GOESMANN, A., MCHARDY, A. C., BARTELS, D., BEKEL, T., CLAUSEN, J., KALINOWSKI, J., LINKE, B., RUPP, O., GIEGERICH, R., & PÜHLER, A. (2003). **GenDB - an open source genome annotation system for prokaryote genomes**. En *Nucleic Acids Research*, vol. 31 de *No. 8*, (pág. 2187–2195). 08 de Abril.

NORMAN, W. E. (2005). **End The Biggest Educational And Intellectual Blunder In History : A $100,000 Challenge To Our Top Educational Leaders**. Scientific Method Publishing. ISBN-10: 0963286668, ISBN-13: 9780963286666. Agosto.

PALMA, Á. A. C. (2011). **Comparación de métodos de agrupamiento: Redes neuronales crecientes y MST-kNN**. Memoria de título profesional de Ingeniero Civil en Informática, Facultad de Ingeniería, Universidad de Santiago de Chile. Santiago, Chile.

RESNIK, P. (1995). **Using information content to evaluate semantic similarity in a taxonomy**. En *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, vol. 1, (pág. 448–453). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ©1995. ISBN: 1-55860-363-8 978-1-558-60363-9; 29 de Noviembre.

ROSS, I., & ROBERT, G. (1996). **R: A Language for Data Analysis and Graphics**.

En *Journal of Computational and Graphical Statistics*, vol. 5 de *3*, (pág. 299–314). DOI: 10.1080/10618600.1996.10474713.

SCHLUETER, S. D., WILKERSON, M. D., HUALA, E., RHEE, S. Y., & BRENDEL, V. (2005). **Community - based gene structure annotation. Trends in Plant Science**. vol. 10 de *No. 1*, (pág. 9–14). 01 de Enero.

SHENOY, M. K., ACHARYA, U. D., & SHET, K. (2012). **A New Similarity measure for taxonomy based on edge counting**. En *International Journal of Web & Semantic Technology*, vol. 3 de *4*, (pág. 23–30). Academy & Industry Research Collaboration Center (AIRCC). ISSN: 0976-2280; DOI: 10.5121/ijwest.2012.3403; Octubre 2012.

SLIMANI, T., YAGHLANEY, B. B., & MELLOULI, K. (2008). **A New Similarity Measure based on Edge Counting**. En *Proceedings of world academy of science, engineering and technology*, vol. 17.

STEIN, L. (2001). **Genome Annotation: From Sequence to Biology**. En *Nature Reviews*, vol. 2, (pág. 493–503). Julio.

UniProtKB (2013). **UniProt Knowledgebase**. Último acceso: Junio 2013.
URL `http://www.uniprot.org/`

VERBANCK, M., LÊ, S., & PAGÈS, J. (2013). **A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data**. En *BMC Bioinformatics*, vol. 14, cap. 3, (p. 42). ISSN: 1471-2105. DOI:10.1186/1471-2105-14-42. 7 de Febrero.

WU, Z., & PALMER, M. (1994). **Verb Semantics And Lexical Selection**. En *ACL '94 Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, (pág. 133–138). Stroudsburg, PA, USA: Association for Computational Linguistics. DOI: 10.3115/981732.981751.

APÉNDICE A. INFORMACIÓN COMPLEMENTARIA

Los apéndices contenidos en la presente sección hacen referencia a información complementaria para la comprensión de tanto las características de la metodología utilizada, como del algoritmo sobre el cual se basa el desarrollo del trabajo.

A.1 CARACTERÍSTICAS DE LAS METODOLOGÍAS UTILIZADAS

Respecto del método científico, diferentes autores han presentado variadas versiones del mismo, o mejor dicho, de las etapas que lo conforman (NORMAN, 2005). Luego de analizarlas, se llega a la conclusión de que la mayoría se describe bajo las siguientes cinco fases, que son finalmente las consideradas para el desarrollo de la tesis:

- **Observación**: Etapa de la metodología denominada así por autores como James K. Feibleman en 1972, semejante a lo que Karl Pearson define en 1892 como "*clasificación cuidadosa y fiel de hechos*"; Conway MacMillan en 1895 como "*reconocimiento y acumulación de hechos*"; Graham Wallas en 1926 como "*preparación*"; Joseph Rossman en 1931 como "*estudio de toda la información disponible*"; James Bryant Conant en 1951 como "*recolección de toda la información relevante*"; Carlo Lastrucci en 1963 como "*estudio de la literatura relacionada y pertinente*"; Irving Copi en 1982 como "*recolección de hechos adicionales*", todas ellas apuntando al concepto al concepto definido por W.I.B. Beveridge en 1950 que dice "*la literatura relevante es revisada críticamente*".

- **Inducción**: Segunda etapa del método científico que será utilizada, denominada así por autores tales como James K. Feibleman en 1972 y similar a la definiciones de "*observación, correlación y consecuencia de hechos*" realizada por Karl Pearson en 1892; "*organización de hechos*" por Conway MacMillan en 1895; "*ponderar toda la evidencia*" hecha por por el Dr. Kenneth Crooks en 1958. Las definiciones anteriores siguen la línea a utilizar en la investigación y descrita en 1950 por W.I.B. Beveridge, que es "*la información obtenida es ordenada y correlacionada, y el problema es definido y dividido en preguntas específicas*".

- **Hipótesis**: Denominada así por autores como James K. Feibleman en 1972 y consecuente con propuestas de esta siguiente etapa para el método científico, como la de Conway MacMillan en 1895 que dice "*enmarcación de la hipótesis*", o como lo denominó John Dewey en 1910, "*sugerencia de una posible solución*"; Graham Wallas en 1926, "*iluminación*"; Joseph Rossman en 1931 "*formulación de todas las soluciones objetivas, análisis crítico de sus ventajas y desventajas y nacimiento de una nueva idea (el invento)*"; el Dr. Kenneth Crooks en 1958 "*hacer la suposición o hipótesis*"; John W. Haefele en 1962, "*hacer el diseño preliminar*"; Irving Copi en 1982, "*formulación de la hipótesis y deducción de consecuencias adicionales*". Todas las definiciones siguiendo la línea de esta investigación, que concuerda con la propuesta por James Bryant Conant en 1951 y que dice "*se formula una hipótesis de trabajo y se escriben deducciones de ella*".

- **Prueba o refutación de la hipótesis**: Denominada así por autores como John W. Haefele en 1962, es la etapa siguiente a la formulación de la hipótesis y concuerda con definiciones de otros autores como Conway MacMillan en 1895 que dice "*la hipótesis es probada y explotada*"; John Dewey en 1910 "*observación adicional y experimentación que conduce a la aceptación o rechazo de la sugerencia*"; James Bryant Conant en 1951 "*las deducciones de la hipótesis sin probadas con ensayos reales y dependiendo del resultado, la hipótesis de trabajo es aceptada, modificada o descartada*"; el Dr. Kenneth Crooks en 1958 "*desafiar la hipótesis*", todas siguiendo la línea que será utilizada, la cual es descrita

por W.I.B. Beveridge en 1950, y que dice "*se idean experimentos para probar primero la hipótesis más probable considerando la mayoría de preguntas cruciales*".

- **Conclusiones**: Algunos autores como Karl Pearson en 1892 definen la etapa como "*autocrítica y toque final que permita una validez en todas las áreas*". Otros como Carlo Lastrucci en 1963 mencionan como "*verificación de la interpretación de información y presentación de los hallazgos en un informe*" o bien como dice la definición más cercana a como será llevada a cabo la fase, descrita por Conway MacMillan en 1895 como "*las conclusiones son escritas, verificadas, aceptadas y aplicadas*".

Respecto de la metodología utilizada para el desarrollo del producto de *Software* R.A.D., es más una compresión (en relación al tiempo) de las etapas de análisis y planificación del proyecto, diseño, construcción, implementación (o integración del desarrollo a un producto mayor) y pruebas, que una metodología que elimine algunas de las etapas, basando su éxito en algunas características específicas, como "tener pocos interlocutores", "tener equipos de trabajo pequeños", "dividir las tareas en otras más pequeñas disminuyendo las características por cada fase de desarrollo para dejarlas a desarrollos posteriores" (lo cual es conocido en la metodología como "*Time Boxing*"), o "utilizar herramientas que permitan la generación de código en base a requerimientos bien definidos" (referenciando al uso de herramientas C.A.S.E.).

A.2 GRAFOS UTILIZADOS POR *MST-KNN*

A.2.1 Árbol de expansión mínima (MST)

El *Árbol de Expansión Mínima* (MST), es un sub-grafo acíclico conectado $Graph_{MST} = (V, E_{MST})$ es decir, todos los vértices están relacionados a al menos otro vértice de manera que la suma de los pesos de las aristas sea mínima (mantener el conjunto de $n - 1$ aristas más

pequeñas que mantengan el grafo conectado). En (INOSTROZA-PONTA, 2008) se propone la implementación de *Prim* para calcular el *MST* de un grafo dada su complejidad computacional de $O(Alog(V))$ y baja carga de memoria.

Algoritmo A.1: Pseudocódigo de la implementación de *Prim* para obtener el *MST* a partir de un grafo.

Data: T: lista de vértices que ya están en el árbol de expansión, y v_1: vértice inicial.
Output: A_{MST}: Aristas asociadas al árbol de expansión mínima del grafo.

```
1  T:= lista vacía;
2  for i := 1 to |V| do
3  |   Q[i] = p_{1,i};
4  |   P[i] = 1;
5  end
6  agregar(T, v_1);
7  while |T| ≠ |V| do
8  |   min = encuentraMin(Q);
9  |   agregar(T, v_min);
10 |   for i := 1 to |V| do
11 |   |   if !pertenece(T, v_i) and (Q[i] > p_{min,i}) then
12 |   |   |   Q[i] = p_{min,i};
13 |   |   |   P[i] = i;
14 |   |   end
15 |   end
16 end
17 for i := 1 to |V| do
18 |   agregar(A_MST, a_{i,P[i]});
19 end
```

El algoritmo A.1 representa a la implementación de *Prim*. En ella, la estructura Q mantiene un registro de las aristas de menor peso (por ende, menor distancia entre los vértices) de aquellos vértices que no se encuentran en el árbol (es decir de $V - T$). La función $encuentraMin(Q)$ entrega como resultado el menor vértice que no haya sido agregado al árbol; la función $agregar(T, v)$ añade el vértice v al conjunto T, y la función $pertenece(T, v)$ retorna verdadero si el vértice v se encuentra en el conjunto T. A partir de un nodo inicial, el algoritmo va iterando y agregando aristas a la estructura de A_{MST} de manera que se mantenga a todos

los vértices conectados a la estructura con su arista de menor peso (o menor distancia). El total de aristas de la estructura es $n-1$, pues son las necesarias para mantener a los n vértices de la estructura inicial conectados.

A.2.2 *k* vecinos más cercanos (kNN)

Los **k** vecinos más cercanos implican la generación de un grafo *kNN*, donde cada vértice tiene a k vértices vecinos cuyas aristas tienen menor peso. Formalmente, una arísta $a_{ij} \in A_{kNN}$ sí y sólo si, el vértice v_j es uno de los k vecinos más cercanos del vértice v_i.

Algoritmo A.2: Pseudocódigo de una posible implementación para obtener el *kNN* a partir de un grafo.

Data: L: lista de vértices.
Output: A_{kNN}: Aristas para formar los sub-grafos de los k vecinos más cercanos.

```
1 for i := 1 to |V| do
2 |   L = cercano(N(i), k);
3 |   for j := 1 to k do
4 |   |   agregar(A_kNN, a_{i,L_j});
5 |   end
6 end
```

El algoritmo A.2, (INOSTROZA-PONTA, 2008) permite obtener el grafo *kNN* a partir de un grafo de aristas ponderadas, con una complejidad computacional de $O(kn^2)$. En él, la función $agregar(A_{kNN}, a_{i,j})$ añade la arista $a_{i,j}$ a la estructura A_{kNN}, y $cercano(N(i), k)$ entrega los k vértices más cercanos de la vecindad de i, representada por $N(i)$.

APÉNDICE B. ANOTACIONES BIOLÓGICAS

Los apéndices contenidos en la presente sección hacen referencia a contenidos complementarios relacionados a las características del conocimiento biológico externo que se desea a incorporar al análisis de experimentos de expresión génica, ya sea a nivel de las características de los elementos que proveen dicha información, como al nivel de cómo ha de relacionarse entre sí dicha información.

B.1 *GENE ONTOLOGY*

B.1.1 Consorcio

La lista de los veinte miembros del consorcio de GO y la descripción de su contribución al proyecto, se presenta a continuación:

1. *Berkeley Bioinformatics Open-source Project*: Desarrollo, uso e integración de ontologías para el análisis de datos biológicos.

2. *British Heart Foundation*: Entrega anotaciones biológicas relacionadas al sistema cardiovascular.

3. *dictyBase*: Base de información del moho *Dictyostelium discoideum*.

4. *EcoliWiki*: Provee información del organismo *Escherichia coli*.

5. *FlyBase*: Fuente de datos de la mosca de fruta *Drosophila melanogaster*.

6. *GeneDB*: Base de datos de variados parásitos protozoarios.

7. *GO Editorial Office*: Establece el contacto con la oficina editorial de GO.

8. *Gramene*: Mantiene información de cereales.

9. *Institute of Genome Sciences, Univ. of Maryland*: Provee datos y herramientas para la investigación genómica en variados sistemas de modelamiento.

10. *InterPro*: Proporciona análisis funcionales de proteinas.

11. *J Craig Venter Institute*: Base de datos de bacterias.

12. *Mouse Genome Informatics*: Mantiene información del ratón *Mus musculus*.

13. *Pombase*: Base de datos de la levadura de fisión *Schizosaccharomyces pombe*.

14. *Rat Genome Database*: Fuente de datos de la rata *Rattus norvegicus*.

15. *Reactome*: Base de conocimiento de procesos biológicos.

16. *Saccharomyces Genome Database*: Base de datos de la levadura *Saccharomyces cerevisiae*, de la cual se utilizó la información que influyó directamente en el trabajo.

17. *The Arabidopsis Information Resource*: Mantiene información de la planta *Arabidopsis thaliana*.

18. *UniProtKB-Gene Ontology Annotation*: Provee anotaciones biológicas manuales y electrónicas de proteinas.

19. *WormBase*: Fuente de datos del nematodo *Caenorhabditis elegans*.

20. *The Zebrafish Information Network*: Referencia fuentes de datos e información de *Danio rerio*.

Además se tiene seis colaboradores, los cuales son los que se listan a continuación:

1. *AgBase*: Sitio *Web* para el análisis funcional de productos génicos de plantas y animales.
2. *AstraZeneca*: Compañía biofarmacéutica relacionada al descubrimiento, desarrollo, manufacturación y venta de medicinas.
3. *Candida Genome Database*: Base de datos del genoma del organismo *Candida*.
4. *Muscle TRAIT*: Mantiene información de transcripciones expresadas del músculo esquelético humano.
5. *Plant-Associated Microbe Gene Ontology*: Colabora con el desarrollo de fuentes de anotaciones para investigadores de la geómica de microbios asociados a plantas.
6. *Tetrahymena Genome Database*: Base de datos del genoma del organismo *Tetrahymena thermophila*.

B.1.2 Estructura de la base de datos de GO

GO provee de forma gratuita todos los elementos necesarios para montar la base de datos en un computador personal para variados motores de bases de datos (como *MySQL*, *OBO* XML, *OWL*, *RDF XML* y *SQL*). Los servicios *Web* que permiten el acceso a GO, lo hacen a una base de datos relacional (del motor *MySQL*) que mantiene, en términos generales, a las ontologías (descritas en la sección 2.2.1), las anotaciones biológicas y los productos génicos.

La base de datos de GO se mantiene a intervalos de tiempo definidos, teniendo en consideración que las anotaciones de *UniProtKB* son omitidas de las bases de datos de actualización semanal y diaria por el tamaño de sus conjuntos de datos. A pesar de lo anterior, se mantienen especies como la *Drosophila* o el *Saccharomyces* por ser autoritarias, gracias a la ayuda de un grupo dedicado a la recopilación y actualización de esos datos.

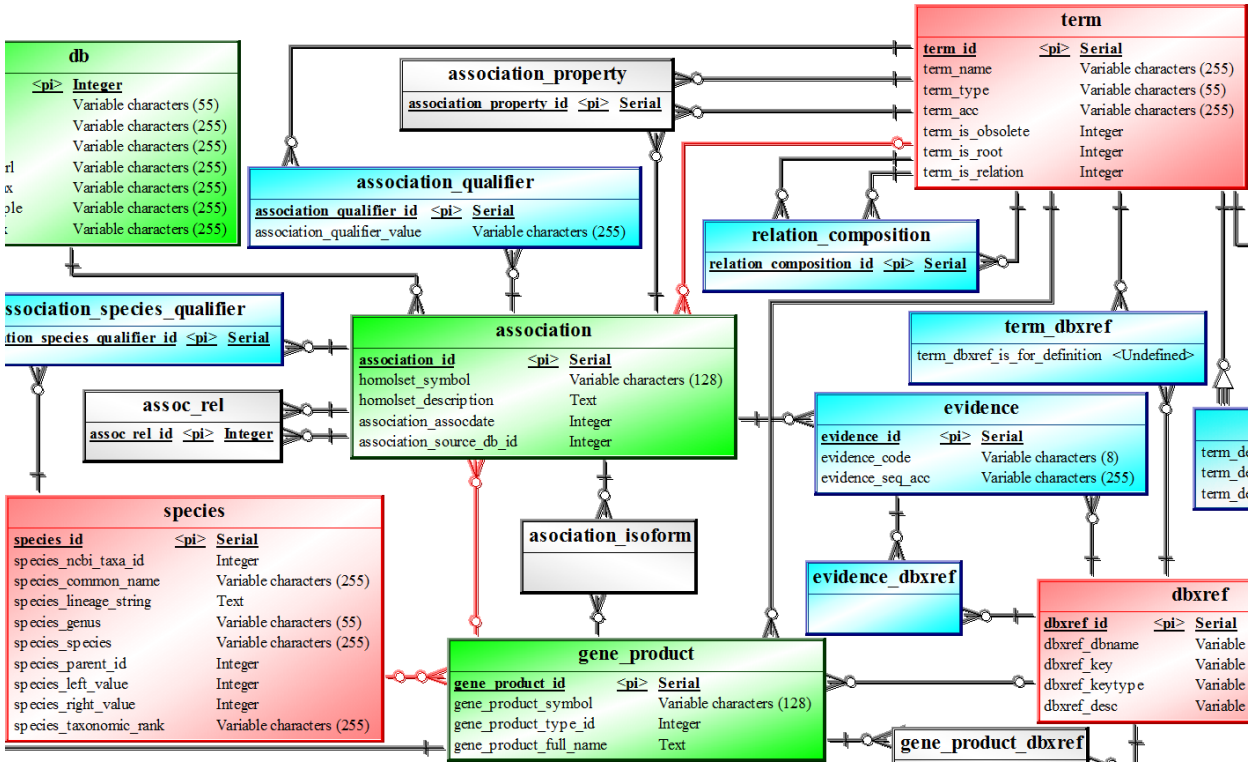

FIGURA B.1: *Extracto del modelo físico de base de datos para Gene Ontology, donde se observan las cuatro principales tablas utilizadas term, association, gene_product y species.*

Actualmente hay una variedad de herramientas que permiten el acceso a la base de datos GO. Un tipo de herramienta funciona a través de la *Web*, no implicando la instalación de componentes relacionados a la base de datos y facilitando por tanto su uso al poseer las funcionalidades de búsqueda, navegación, descarga e incluso análisis y procesamiento de los datos, motivo por el cual son ampliamente usadas y un campo de desarrollo constante. Otra opción es instalar un *Software* en un computador personal, y además la base de datos o archivos necesarios para su correcto funcionamiento. Dado que las herramientas *Web* tienen como requisito el estar conectado a una red, y depender de la disponibilidad del proveedor del servicio, se mantienen espejos de la base de datos accesibles a través de clientes *MySQL*. Si bien GO permite la realización de consultas directas en lenguaje *SQL* (por ejemplo, con la herramienta *GOOSE*), soporta además consultas en otros lenguajes como *Perl* y *JAVA*.

En la figura B.1 se presenta un extracto del modelo conceptual de datos de GO, el

cual actualmente tiene un total de 44 entidades. Las entidades de color rojo representan a un concepto fundamental del modelo (posee una gran cantidad de relaciones, con respecto a las demás entidades), las de color verde representan una concepto concreto, las entidades de color cian son creadas para mantener relaciones entre otras entidades (generalmente por la formación de relaciones "$N : N$"), y las de color gris entidades que no contienen ningún registro asociado (de las cuales hay un total de 21). Del modelo completo, fueron utilizadas cuatro entidades

- ***species***: Mantiene información taxonómica para un organismo. Sus atributos son, un identificador (*id*), un identificador consecuente a la taxonomía de *NCBI* (*ncbi_taxa_id*), un nombre no científico (*common_name*), una lista desnormalizada de nombres de taxón como texto (*lineage_string*), el género o nombre científico del taxón (*genus*), nombre de la especie (*species*), taxón padre de la jerarquía (*parent_id*), valor derecho e izquierdo visto como un modelo anidado (*left_value* y *rigth_value*) y clasificación de la especie (*taxonomic_rank*).

- ***gene_product***: Mantiene información que representa a un gen, o un producto génico a nivel de especie. Sus atributos son, un identificador (*id*), una etiqueta para el producto génico (*symbol*) y un símbolo usado típicamente como etiqueta concisa (*full_name*).

- ***association***: Mantiene una relación entre un producto génico y un término biológico con más de una evidencia. Los atributos son, un identificador (*id*), un número que si es distinto de cero indica que el producto del gen no tiene un rol definido por el término biológico (*is_not*) y la fecha en que se llevó a cabo el último chequeo de la asociación de las bases de datos de los proveedores, en formato "YYYYMMDD" (*assocdate*).

- ***term***: Mantiene las unidades de representación fundamentales de las ontologías (nodos del grafo de la ontología), y también relaciones, como *is_a* y *part_of*. Los atributos son, un identificador (*id*), una etiqueta textual para el término (*name*), la ontología o espacio de nombres al cual pertenece el término (*term_type*), el identificador único para el término

(*acc*), un número que de ser cero indica que el término no está obsoleto (*is_obsolete*), un entero que de tener valor uno indica que el nodo es la raíz en el grafo de ontología (*is_root*) y un número que de tener valor uno indica que el término representa a una relación (*is_relation*).

B.1.3 *OBO Flat File Format*

Con el objetivo de no establecer de forma manual las relaciones entre los términos (existe una tabla en la base de datos de GO que relaciona dos términos entre sí), se hace uso del archivo de texto plano OBO, el cual está implementado para ser tratado con el *Software OBO-Edit* (desarrollado en *JAVA*), el cual a partir del archivo OBO genera una estructura de datos DAG de las anotaciones biológicas identificando si un término es ancestro o descendiente de otro.

El archivo OBO se encuentra disponible en dos formatos, *OBO format* y *GO slims*, siendo el segundo una versión resumida del contenido del primero. En particular, se hizo uso de *OBO format* en su versión *v1.2*, el cual es de actualización diaria y posee todo el conjunto de ontologías con los términos y sus relaciones en un formato legible, de fácil análisis y extensible. El archivo ignora líneas en blanco y se basa en una estructura dividida en un encabezado (conformado por etiquetas y valores para ellas) y el contenido mismo expuesto en estrofas, como se grafica a continuación:

```
<encabezado>
<estrofa>
<estrofa>
...
```

El encabezado permite describir al archivo a través de características como la versión a la que corresponde (etiqueta requerida), la versión de la ontología, la fecha y hora de generación, el nombre del usuario que guardó por última el archivo, el nombre del *Software* que generó el archivo, descripciones de los subconjuntos de términos, entre otras etiquetas opcionales que

permiten diferenciarlo. A continuación se presenta un extracto del encabezado del archivo OBO utilizado.

```
format-version: 1.2
data-version: 2013-05-14
date: 13:05:2013 04:06
saved-by: al
auto-generated-by: TermGenie 1.0
subsetdef: goslim_plant "Plant GO slim"
...
subsetdef: goslim_yeast "Yeast GO slim"
...
subsetdef: virus_checked "Viral overhaul terms"
synonymtypedef: systematic_synonym "Systematic synonym" EXACT
default-namespace: gene_ontology
remark: cvs version: $Revision: 8538 $
ontology: go
```

Por otro lado, las estrofas describen a los términos (y otros tipos de objetos identificados por las etiquetas *Typedef* e *Instance*) contenidos en el archivo a través de pares de identificadores y valores de diferentes características (según sea el tipo de objeto), en base a la siguiente estructura:

```
[Etiqueta del tipo de objeto]
<ID de caract.>:<valor de caract.>
<ID de caract.>:<valor de caract.>
...

[Etiqueta del tipo de objeto]
<ID de caract.>:<valor de caract.>
<ID de caract.>:<valor de caract.>
...
```

Dentro de cada estrofa hay una serie de identificadores que pueden utilizarse, y hacen referencia a diferentes características del objeto, como por ejemplo (para el caso del objeto "*Term*") el nombre del término, su identificador único en la base de datos, definición, comentarios asociados, sinónimo, equivalencia del término con respecto a la intersección de otros términos, creador, fecha y hora de creación, entre otras, como se presenta en el siguiente ejemplo:

```
[Term]
id: GO:0000003
name: reproduction
namespace: biological_process
alt_id: GO:0019952
```

```
alt_id: GO:0050876
def: "The production by an organism of new individuals that contain some portion of their
    genetic material inherited from that organism." [GOC:go_curators, GOC:isa_complete, ISBN
    :0198506732]
subset: goslim_generic
subset: goslim_pir
subset: goslim_plant
subset: gosubset_prok
synonym: "reproductive physiological process" EXACT []
xref: Wikipedia:Reproduction
is_a: GO:0008150 ! biological_process
```

Los comentarios dentro del archivo (ignorados por los analizadores del contenido) corresponden al texto que se encuentre entre el caracter "!" y "\", sin ser permitida una línea con "\" dentro de un comentario. El archivo *OBO* contiene a pares etiqueta-valor, donde la etiqueta es una cadena de caracteres que puede o no estar predefinida (un analizador no arroja un error de no reconocerla, lo que permite agregar nuevas etiquetas), y el valor que es una cadena de caracteres que puede requerir de un análisis diferente dependiendo la etiqueta a la que esté asociado. En general los pares etiqueta-valor se conforman de una línea a menos que se haga uso de caracteres de escape, que son aquellos que poseen un significado especial para el analizador (para identificar, por ejemplo, una nueva línea, tabulación, paréntesis, comillas, etc. siempre antecedidos por un *backslash*). A continuación se presenta el formato común de un par etiqueta-valor:

```
<etiqueta>: <valor> {<nombre>=<valor>, <nombre>=<valor>, ...} ! <comentario>
<etiqueta>: <valor> {<nombre>=<valor>, <nombre>=<valor>, ...} ! <comentario>
...
```

Del ejemplo anterior, los elementos contenidos entre llaves ("{ }") corresponden al uso de modificadores. Introducidos en la versión 1.2 de *OBO*, los modificadores permiten agregar nuevas características a etiquetas pre-definidas o existentes, y dependiendo del analizador, la información que contengan puede ignorarse o decodificarse, por lo que su contenido debe ser opcional o experimental.

B.2 DISTANCIA SEMÁNTICA DE TÉRMINOS BIOLÓGICOS

B.2.1 Intuiciones y supuestos de la medida de similitud de *Lin*

La medida de similitud propuesta por (LIN, 1998), se basa en las siguientes intuiciones y supuestos para el cálculo de la similitud entre términos de una ontología, lo que permite su generalización a variadas áreas de la investigación:

- ***Intuición N°1***: La similitud entre el elemento A y el elemento B se relaciona a lo que posean en común, de manera que mientras más cosas compartan A y B, más similares han de ser entre si.

- ***Supuesto N°1***: Considerando la *Intuición N°1*, el cálculo de lo que poseen en común los elementos A y B, es a través de:

  - $Comun_{A,B} = CI\left(Com\left(A,B\right)\right)$

  Donde Com representa lo que posee en común A con B.

- ***Intuición N°2***: La similitud entre el elemento A y el elemento B se relaciona a lo que no posean en común, de manera que mientras más diferencias posean A y B, menos similares han de ser entre sí.

- ***Supuesto N°2***: Considerando la *Intuición N°2*, el cálculo de lo que no poseen en común los elementos A y B, es a través de:

  - $Different_{A,B} = CI\left(Desc\left(A,B\right)\right) - Comun_{A,B}$

  Donde $Desc$ representa a la descripción de lo son A y B.

- ***Intuición N°3***: Cuando el elemento A y el elemento B son idénticos, poseen el máximo valor de similitud, sin importar lo que posean en común.

- ***Supuesto N°3***: Dado que se puede cuantificar lo que poseen en común los elementos A y B, se establece una medida de similitud que considere ese valor a través del cálculo de:

  - $Sim\left(A,B\right) = f\left(Comun_{A,B}, CI\left(Desc\left(A,B\right)\right)\right)$

  Donde f tiene un dominio de $\{(x,y)\,|x \geq 0, y > 0, y \geq x\}$.

- ***Supuesto N°4***: Cuando los elementos A y B son idénticos (*Intuición N°3*), al considerar el valor uno como el valor máximo para la similitud, la función f debe cumplir con:

  - $\forall x > 0, f\left(x,x\right) = 1$

- ***Supuesto N°5***: El último supuesto a considerar, corresponde a:

  - $\forall x_1 \leq y_1, x_2 \leq y_2 : f\left(x_1 + x_2, y_1 + y_2\right) = \dfrac{y_1}{y_1 + y_2} f\left(x_1, y_1\right) + \dfrac{y_2}{y_1 + y_2} f\left(x_2, y_2\right)$

APÉNDICE C. RESULTADOS OBTENIDOS

El presente apéndice tiene por objetivo complementar los resultados obtenidos dados a conocer en el capítulo cuatro, en caso de que se desee un mayor detalle respecto de las clasificaciones obtenidas para las parametrizaciones, con respecto a los índices de validación para los experimentos realizados.

C.1 CLASIFICACIÓN DE ENFOQUES DE DISTANCIA SEMÁNTICA

Respecto de la clasificación para los enfoques de distancia semántica (figuras C.1, C.2 y C.3), se muestran los valores para el 5 %, 10 %, 25 % y 50 % mejor clasificado.

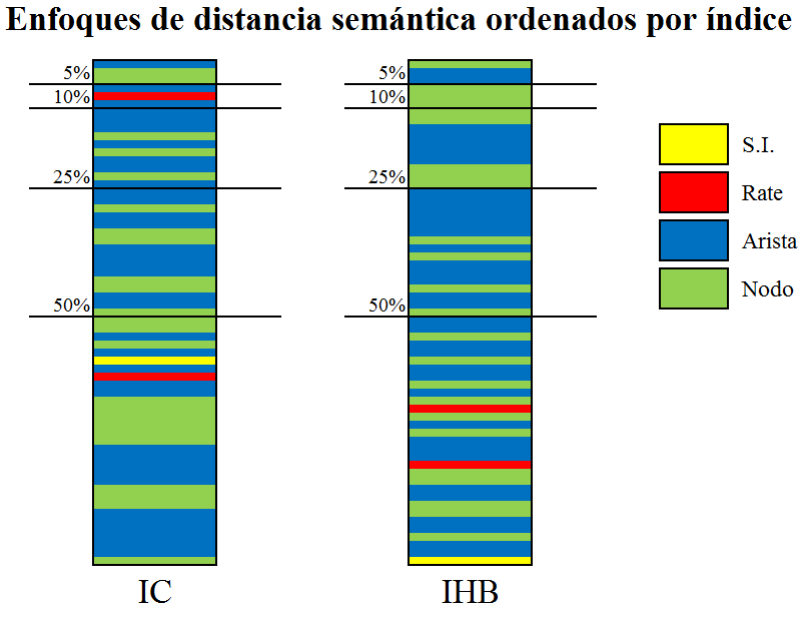

FIGURA C.1: *Clasificación de los enfoques de similitud semántica (S.I., Nodos, Aristas y Tasa) clasificados por IC e IHB, para* $Experimento_{1-\rho}$.

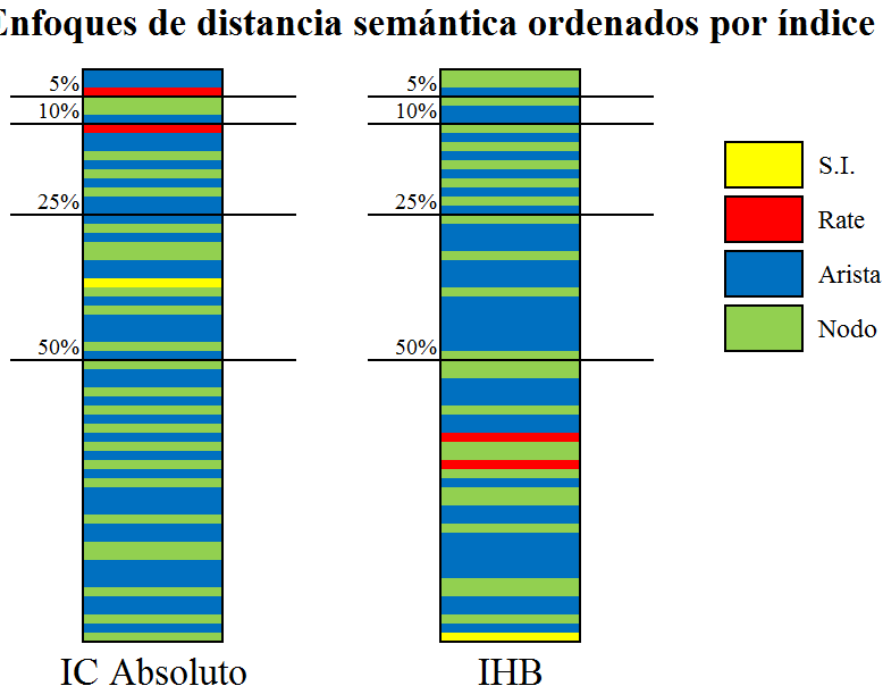

FIGURA C.2: *Clasificación de los enfoques de similitud semántica (S.I., Nodos, Aristas y Tasa) clasificados por IC Absoluto e IHB, para* $Experimento_{1-|\rho|}$.

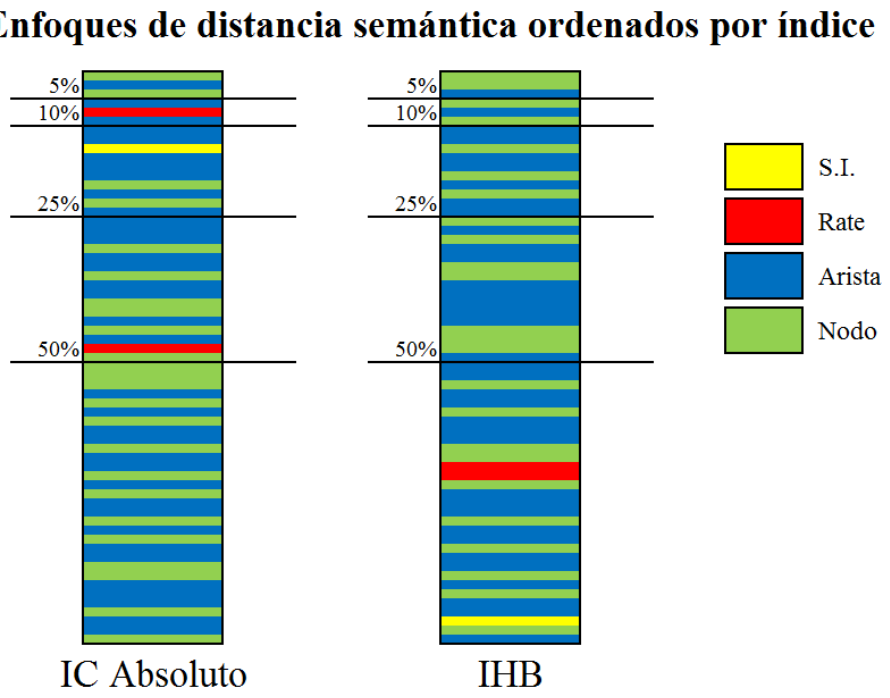

FIGURA C.3: *Clasificación de los enfoques de similitud semántica (S.I., Nodos, Aristas y Tasa) clasificados por IC Absoluto e IHB, para* $Experimento_{\rho+\rho}$.

C.2 CLASIFICACIÓN DE DISTANCIA SEMÁNTICA PARA ANOTACIONES BIOLÓGICAS

Respecto de la clasificación para las medidas de distancia semántica de *Wu-Palmer* (WU & PALMER, 1994), T.B.K. (SLIMANI et al., 2008), *Leacock-Chodorow* (LEACOCK & CHODOROW, 1998), *Jiang-Conrath* (JIANG & CONRATH, 1997) y *Lin* (LIN, 1998), se muestran los valores para el 5 %, 10 %, 25 % y 50 % mejor clasificado (además de mostrar las medidas peor clasificadas) en las figuras C.4 (con respecto a los índices *IC* e *IHB*), C.5 y C.6 (con respecto a los índices *IC Absoluto* e *IHB*).

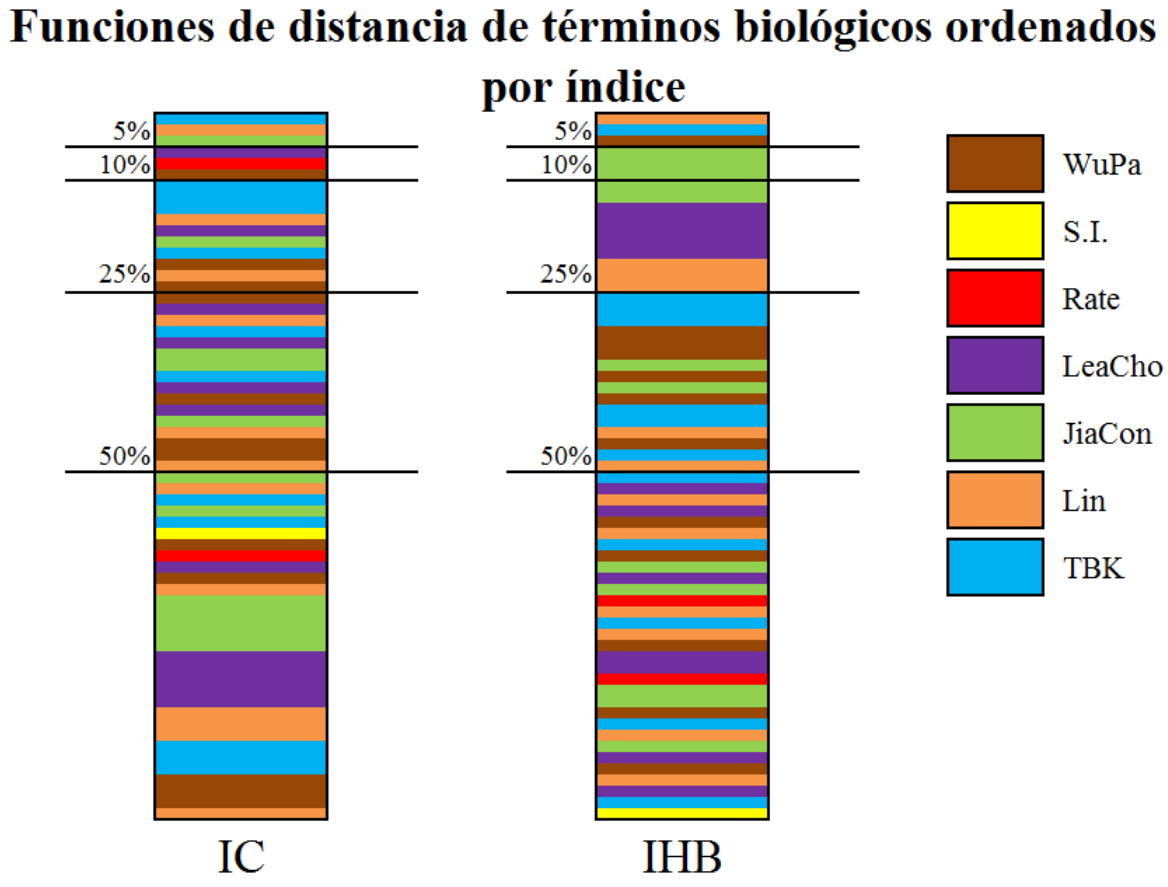

FIGURA C.4: *Clasificación de medidas de distancia entre términos biológicos (S.I., Rate, $Dist_{wp}$, $Dist_{tbk}$, $Dist_{lc}$, $Dist_{jc_{norm}}$ y $Dist_{lin}$) clasificados por IC e IHB, para $Experimento_{1-\rho}$.*

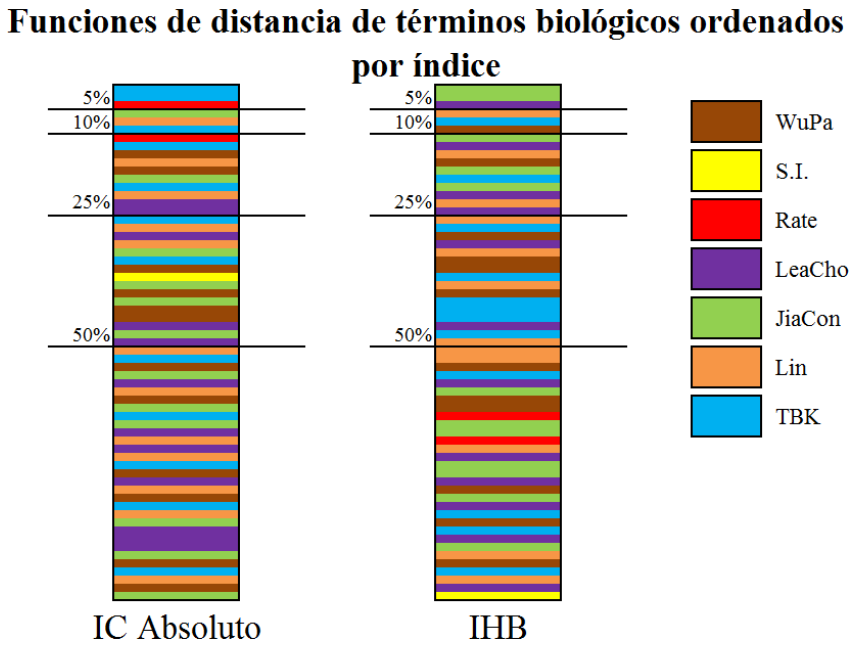

FIGURA C.5: *Clasificación de medidas de distancia entre términos biológicos (S.I., Rate, $Dist_{wp}$, $Dist_{tbk}$, $Dist_{lc}$, $Dist_{jc_{norm}}$ y $Dist_{lin}$) clasificados por IC Absoluto e IHB, para $Experimento_{1-|\rho|}$.*

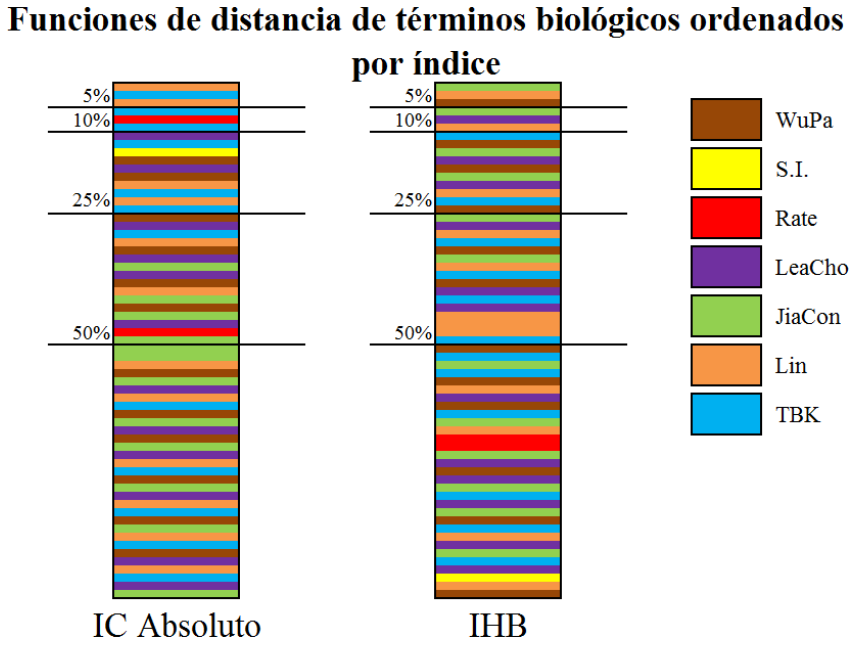

FIGURA C.6: *Clasificación de medidas de distancia entre términos biológicos (S.I., Rate, $Dist_{wp}$, $Dist_{tbk}$, $Dist_{lc}$, $Dist_{jc_{norm}}$ y $Dist_{lin}$) clasificados por IC Absoluto e IHB, para $Experimento_{\rho+\rho}$.*

C.3 CLASIFICACIÓN DE DISTANCIA DE PERFILES FUNCIONALES

Respecto de la clasificación para las medidas de distancia de perfiles funcionales (*Min*, *Max*, *Aver*, *Rate*, *PDP*, *MDP* y *Match*), se muestran los valores para el 5 %, 10 %, 25 % y 50 % mejor clasificado (además de mostrar las medidas peor clasificadas) en las figuras C.7 (con respecto a los índices *IC* e *IHB*), C.8 y C.9 (con respecto a los índices *IC Absoluto* e *IHB*).

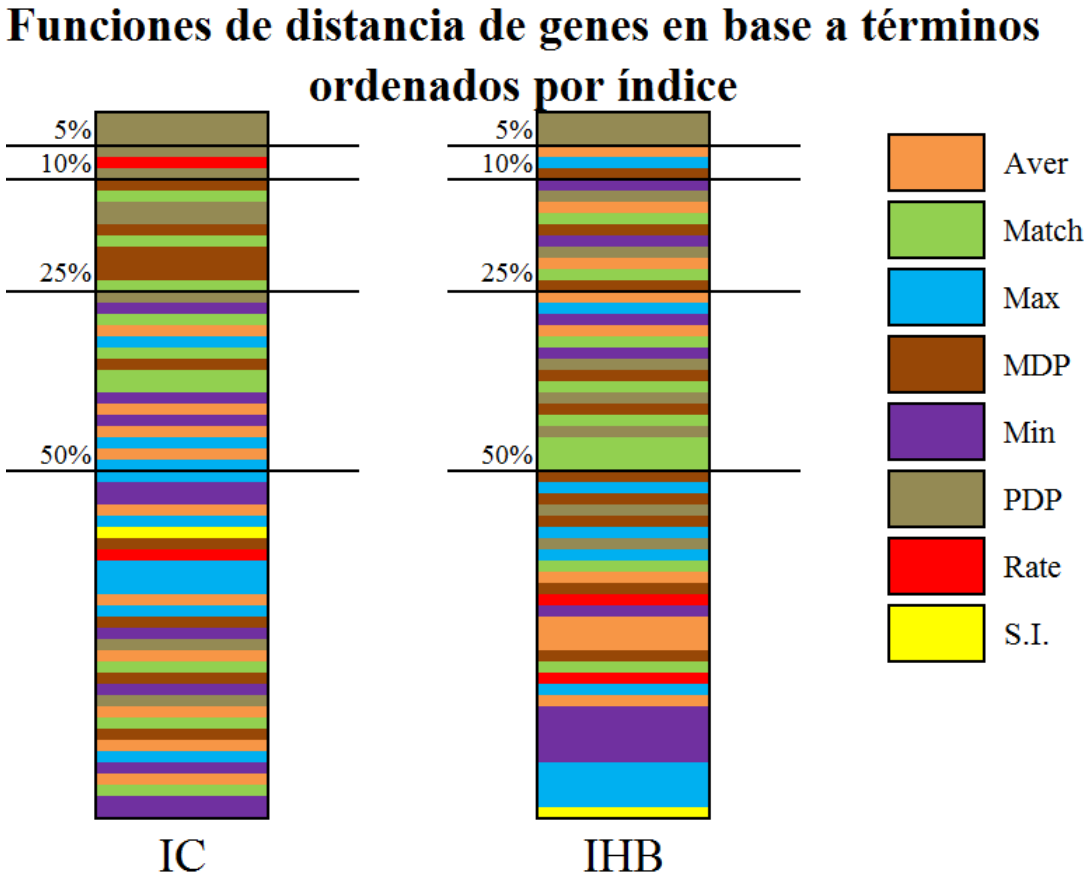

FIGURA C.7: *Clasificación de medidas de distancia entre genes en base al conocimiento biológico (S.I., Rate, Min, Max, Aver, PDP, MDP y Match) clasificados por IC e IHB, para* $Experimento_{1-\rho}$.

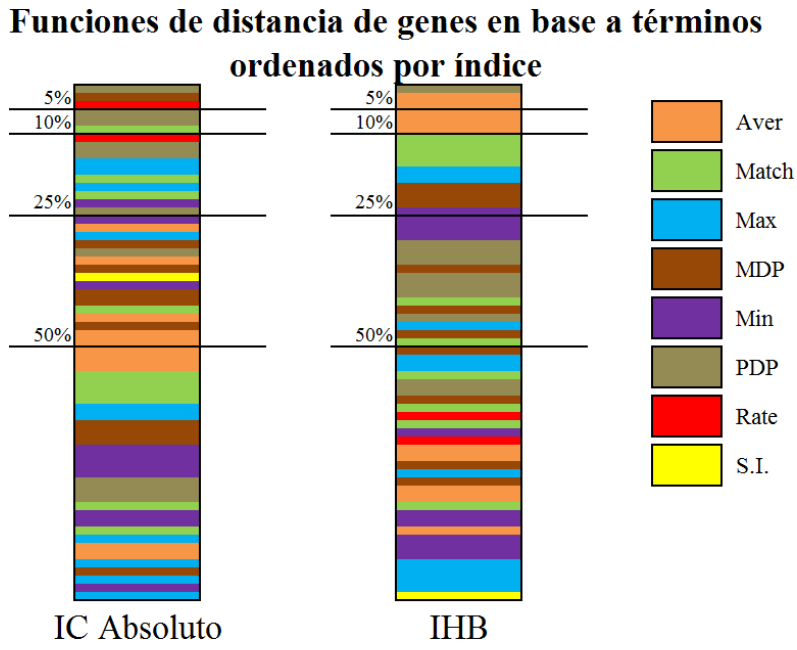

FIGURA C.8: *Clasificación de medidas de distancia entre genes en base al conocimiento biológico (S.I., Rate, Min, Max, Aver, PDP, MDP y Match) clasificados por IC Absoluto e IHB, para* $Experimento_{1-|\rho|}$.

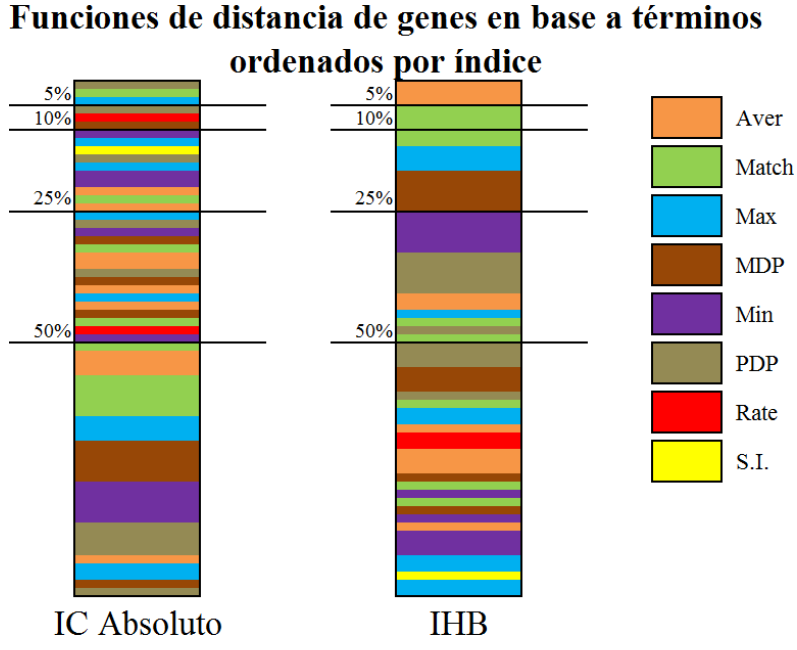

FIGURA C.9: *Clasificación de medidas de distancia entre genes en base al conocimiento biológico (S.I., Rate, Min, Max, Aver, PDP, MDP y Match) clasificados por IC Absoluto e IHB, para* $Experimento_{\rho+\rho}$.

C.4 CLASIFICACIÓN DE COMBINACIÓN DE MATRIZ DE PERFIL FUNCIONAL Y DE EXPRESIÓN

Respecto de la clasificación para las funciones de combinación de matrices de perfiles funcionales con las de expresión génica ($\alpha = 0,5$ y *Eucli*), se muestra los valores para el 5 %, 10 %, 25 % y 50 % mejor clasificado (además de mostrar las medidas peor clasificadas) en las figuras C.10 (con respecto a los índices *IC* e *IHB*), C.11 y C.12 (con respecto a los índices *IC Absoluto* e *IHB*).

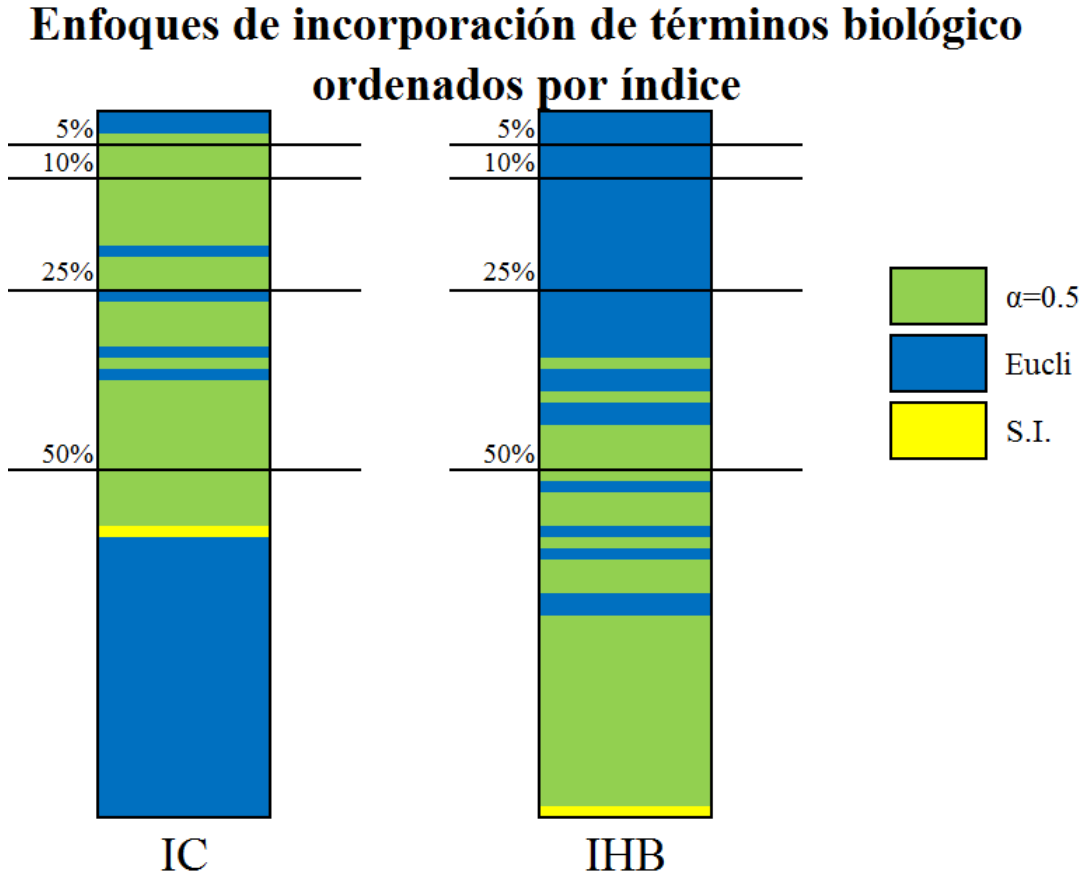

FIGURA C.10: *Clasificación de funciones de incorporación de conocimiento biológico (S.I., $\alpha = 0,5$ y Eucli) clasificados por IC e IHB, para experimento* $Experimento_{1-\rho}$.

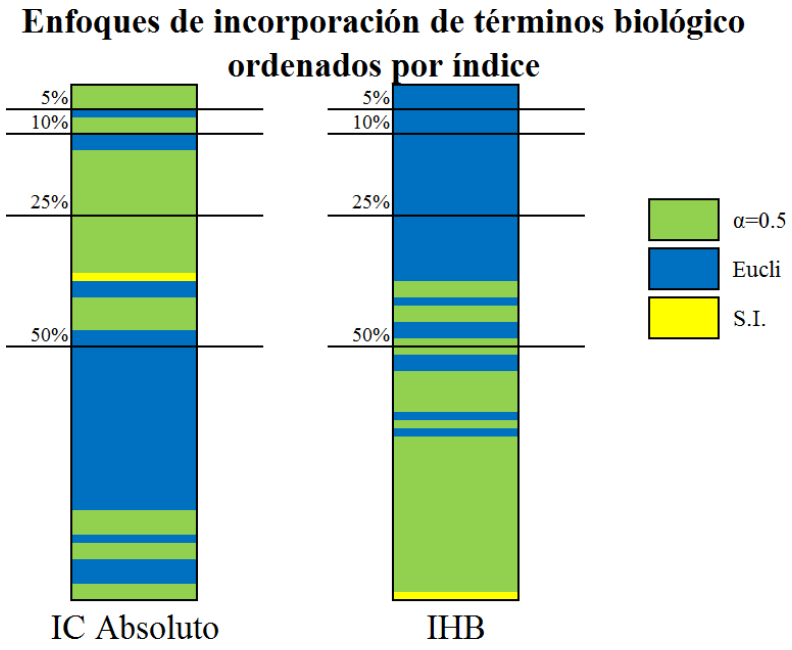

FIGURA C.11: *Clasificación de funciones de incorporación de conocimiento biológico (S.I., $\alpha = 0,5$ y Eucli) clasificados por IC Absoluto e IHB, para experimento* $Experimento_{1-|\rho|}$.

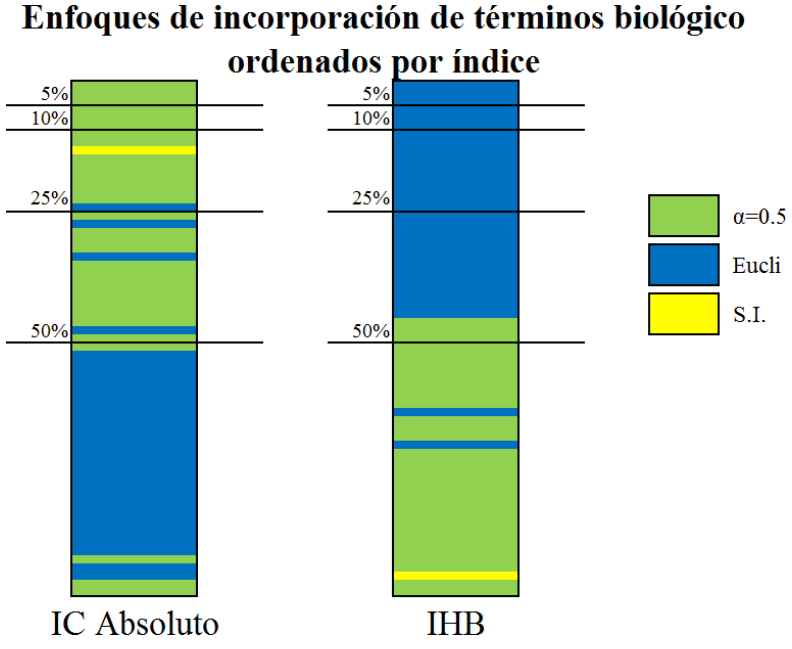

FIGURA C.12: *Clasificación de funciones de incorporación de conocimiento biológico (S.I., $\alpha = 0,5$ y Eucli) clasificados por IC Absoluto e IHB, para experimento* $Experimento_{\rho+\rho}$.