UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"APLICACIÓN DE TÉCNICAS DE COMPUTACIÓN SUAVE A LA CLASIFICACIÓN DE PROTEÍNAS EN *Saccharomyces cerevisiae*"

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A:

ANA MARÍA ESCALANTE GONZALBO

DIRECTOR DE LA TESIS: DR. ANGEL FERNANDO KURI MORALES

MÉXICO, D.F. 2006.

¿Qué sería de la luna sin noche, de la lluvia sin agua, de las horas sin días,
de la brisa sin viento?

¿Qué sería de los cauces sin ríos, de las venas sin sangre, de la mueca sin rostro,
de los versos sin letras,
del esfuerzo sin meta,
de la voz sin aliento?

¿Qué sería de los versos sin sangre, de las letras sin cauces, de la brisa sin rostro,
de las horas sin viento,
de los ríos sin aliento,
de las voces sin días,
de los días sin noche,
de la noche sin vida,
de la vida sin sueños…

Gerardo y Leonardo,

Hechiceros de la vida y maestros de la risa,
artesanos de los sueños y amos de la fantasía,
ráfaga intempestiva,
corriente caudalosa,
explosión repentina
y golpe seco.

Noche, día, agua, viento,
letras, río, sangre, versos,
rostro, meta, voz, aliento,
llanto, risa, vida y sueños.

AGRADECIMIENTOS

En primer lugar quiero hacer patente mi profundo agradecimiento al Dr. Angel Kuri Morales, director de este trabajo, por su tenacidad, tolerancia y paciencia a lo largo de muchos años y sin cuyo apoyo, entusiasmo y confianza, nunca hubiera podido llegar a este punto. Desde la propuesta del problema y el método de análisis, hasta el seguimiento de los avances, la aportación de ideas y sugerencias durante todo el proceso y la revisión de los manuscritos, Angel estuvo presente en todas las etapas de este trabajo. Gracias por ser mucho más que un director de tesis.

También quiero agradecer a Gerardo Coello, cómplice, compañero, asesor, consejero y maestro, sin cuya presencia y apoyo ya no podría entender la vida. Siempre con comentarios acertados, juicios implacables y sugerencias oportunas, tiene la capacidad de ayudarme a sacar lo mejor de mi misma y resolver los problemas que parecían casi imposibles. Con sus conocimientos e ingenio me ayudó a resolver varios de los problemas de validación a los que me enfrenté durante el desarrollo de este trabajo y me ayudó en el desarrollo de algoritmos y programas.

A la Dra. Ana Lilia Laureano y a los Doctores Francisco Cervantes, Roberto Coria y Nicolás Kemper, quiero agradecerles que hayan aceptado participar como sinodales de este trabajo y hayan dedicado su tiempo a leer mi manuscrito y hacerme sugerencias y comentarios oportunos y enriquecedores.

A José Galavíz, Carlos Galindo, Carlos Guadarrama, Lucía Castellanos e Iván Mejía, compañeros del seminario de Bioinformática, por sus comentarios y aportaciones a este trabajo.

De manera especial, quiero agradecer al Dr. Jesús Adolfo García-Saínz, director del Instituto de Fisiología Celular, institución en la que trabajo desde hace más de diez años, por su apoyo decidido y generoso, enfocado a favorecer tanto mi desarrollo profesional como académico.

En el terreno personal, quiero agradecer a Francisco Pérez, Ivette Rosas, Juan Manuel Barbosa y Sergio Rojas, amigos y compañeros de trabajo, que con su profesionalismo y dedicación constantes me permitieron disponer del tiempo necesario para llevar a cabo este trabajo.

Aún sabiendo que un agradecimiento no es ni necesario ni suficiente, no puedo dejar de mencionar a Pilar Gonzalbo, quien durante todos estos años ha sido mucho más que una abuela para mis hijos y sin cuya ayuda desbocada y apoyo incondicional no podría ni imaginar haber continuado mi desarrollo profesional. Gracias por estar siempre ahí cuando te necesito.

Por último, no quiero dejar de mencionar a la Sra. María del Lourdes González, secretaria del Posgrado en Ciencias e Ingeniería de la Computación en el IIMAS, quien siempre ha tenido la mejor disposición, no solamente para proporcionarme información e indicarme los trámites a seguir, sino haciendo todo lo que está a su alcance para ayudarme y facilitarme cada paso del proceso.

INDICE DE CONTENIDO

RESUMEN

La información necesaria para que los organismos vivos realicen la totalidad de sus funciones biológicas se encuentra almacenada en las moléculas de DNA (Acido Desoxirribonucléico) que constituyen su genoma. Mucho se ha descubierto en los últimos años en relación a las reglas que gobiernan tanto la expresión de la información almacenada en el DNA, como algunos de los mecanismos asociados con la regulación de la expresión génica. Sin embargo, aún falta mucho por saber en lo que respecta a la relación entre secuencia y estructura y función de las proteínas, que se sintetizan a partir de las moléculas de DNA. Gracias a las técnicas de secuenciación se cuenta ya con un enorme volumen de información de secuencias de proteínas para muchos organismos, sin embargo, se desconoce aún la estructura y función de la gran mayoría de éstas.

Si partimos del hecho de que la secuencia de aminoácidos determina tanto la estructura como la función de las proteínas, entonces debería ser posible identificar grupos de proteínas semejantes y de esta forma poder inferir la estructura y función de las proteínas desconocidas, a partir de las ya identificadas que pertenezcan al mismo grupo. Los esfuerzos que se han hecho en este terreno y que en la mayoría de los casos involucran algún tipo de alineamiento de secuencias, no han obtenido resultados totalmente satisfactorios.

En este trabajo, proponemos un nuevo método de clasificación de secuencias de proteínas, mediante la utilización de un procedimiento de clasificación no supervisado, conocido como mapas autoorganizados (SOMs), o redes de Kohonen, que pretende obtener una clasificación no sesgada de las secuencias. Es decir, permitir que las proteínas se agrupen entre sí en términos de su similitud de secuencia, sin establecer limitaciones o criterios a priori que puedan dirigir la clasificación en algún sentido.

Con la finalidad de probar este método, utilizamos la totalidad de secuencias de proteínas de la levadura *Saccharomyces cerevisiae*, que es una de las especies más estudiadas y de la que se cuenta con gran cantidad de información.

Cada proteína se representó como un vector de características en el que se buscó incluir la mayor cantidad de información de la secuencia, en términos de frecuencia y posición de los aminoácidos. La secuencia se organizó en arreglos de una, dos, tres y cuatro dimensiones. Estos vectores se utilizaron como entrada para entrenar el mapa de Kohonen.

El número óptimo de clases se determinó utilizando un criterio fundamentado en el análisis de tendencias en la entropía del algoritmo fuzzy-c means y se demostró la validez de utilizar este criterio mediante comparaciones de los conjuntos obtenidos tanto con este método como con los mapas autoorganizados.

Posteriormente, se demostró estadísticamente, que los conjuntos obtenidos mediante los SOMs no responden a criterios triviales como longitud de la secuencia o frecuencia de aminoácidos y, por último, se llevó a cabo la validación de la clasificación desde el punto de vista biológico, comparando los resultados obtenidos con una base de datos de función de las proteínas.

La base de datos que se utilizó para la validación está desarrollada por el consorcio Gene Ontology (GO), que establece relaciones de carácter funcional entre las proteínas de varias especies y se encuentra particularmente completa para el caso de la levadura.

Con la finalidad de calificar el ajuste de nuestros conjuntos a categorías funcionales de GO, desarrollamos un índice de calidad Q. Este nos permite determinar el grado de correspondencia de los conjuntos, con respecto a categorías de GO. Generamos conjuntos aleatorios de proteínas (de tamaños semejantes a nuestros conjuntos) y comparamos estadísticamente el ajuste de estos conjuntos y el de los que nosotros obtuvimos mediante SOMs. La estadística mostró que sí hay diferencias significativas entre el ajuste de los conjuntos aleatorios y los que nosotros generamos, notándose mayores valores del índice Q para nuestros conjuntos.

Se observó una tendencia clara de nuestros conjuntos a ocupar categorías funcionales específicas y se hizo una descripción general en términos funcionales de cada uno de ellos.

Por último, se hizo un ensayo de utilizar nuestra red entrenada como un sistema clasificador y se encontró que más del 80% de las proteínas clasificadas estaban contenidas en nuestra descripción de las clases. También se encontró que nuestras descripciones resultaban demasiado generales. No permitían discriminar de manera categórica entre proteínas pertenecientes a distintas clases.

De acuerdo con estos resultados, pudimos concluir que el método de clasificación propuesto tiene potencialidades interesantes en el campo de la clasificación de proteínas, pero que sería deseable poder realizar subclasificaciones adicionales para poder llegar a definiciones más específicas y excluyentes de cada una de las clases.

Más allá de la posibilidad de asignar estructura y función a proteínas de función desconocida, el método podría utilizarse también como punto de partida para la identificación de patrones o motivos característicos de cada clase, que pudieran proporcionar alguna pista sobre las claves ocultas en la determinación de la estructura y función de las proteínas. Utilizando algoritmos de búsqueda de patrones complejos en cada una de las clases, podrían identificarse motivos característicos de cada una de ellas y ausentes en las demás, que proporcionarían la clave de su pertenencia a dicha clase y por extensión, de su estructura y función características.

El trabajo que aquí se presenta, representa una aportación interesante al campo de la bioinformática. Es un esfuerzo por encontrar una clasificación no sesgada de las proteínas que permita que las características distintivas que pudieran estar ocultas en la secuencia, emerjan de manera natural. Abre un extenso terreno de exploración tanto en el campo de la clasificación, como en el de búsqueda de patrones, que podría ayudar a resolver la clave de la relación que existe entre secuencia y estructura y función de las proteínas.

1

INTRODUCCIÓN

En los últimos 10 años, debido al gran desarrollo de la biología molecular y al perfeccionamiento de las técnicas de secuenciación de genes, ha habido un crecimiento explosivo en la cantidad de datos de secuencias de proteínas de muchos organismos. Sin embargo, las técnicas para analizar y organizar esta información no se han desarrollado a la misma velocidad. Los biólogos moleculares se enfrentan ahora con el gran reto de organizar estos enormes volúmenes de datos, tratando al mismo tiempo de extraer la mayor cantidad de información que permita desentrañar las reglas que gobiernan el funcionamiento y la estructura de estas moléculas indispensables para la vida.

Las ciencias de la computación han empezado a incorporarse a la biología, aportando herramientas y algoritmos que ayudan en el análisis y manejo de este universo de nuevos datos. Sin embargo la mayoría de los métodos aplicados en concreto al problema de la clasificación de proteínas, están basados en aproximaciones apriorísticas, en las que los criterios del investigador sesgan los resultados obtenidos, y aunque en un principio se consiguieron avances importantes, en la actualidad el campo parece estar atrapado en su propio paradigma reduccionista. En el presente trabajo pretendemos dar un nuevo enfoque al problema de la clasificación de proteínas, que mediante la utilización de técnicas de computación suave y aprendizaje no supervisado, permita que, si existen relaciones de mayor orden que gobiernan la estructura y funcionamiento de las proteínas, dicha clasificación emerja de manera natural, sin establecer criterios *a priori* que constriñan la clasificación en alguna dirección preestablecida.

Proponemos un método novedoso para la representación de las secuencias de proteínas, que integra elementos tanto de la composición como de la estructura de las secuencias y la utilización de mapas autoorganizados para obtener una clasificación de las proteínas de la levadura (*Saccharomyces cerevisiae*) a partir de dicha representación. A continuación hacemos una validación de los grupos obtenidos utilizando como referencia la base de datos Gene Ontology: una estructura jerárquica funcional de anotaciones a productos génicos, que es particularmente rica para el caso de la especie antes mencionada.

1.1 Planteamiento del problema

1.1.1 De los datos al conocimiento

El **conocimiento** puede definirse como un conjunto de información estructurada y organizada de forma que pueda ser utilizada. Si bien el conocimiento tiene su base en la acumulación de **datos**, los datos en si mismos no constituyen conocimiento. Los datos pueden ser considerados como la materia prima de la **información**, son un conjunto de hechos u observaciones. Un grupo de datos estructurados y organizados constituyen información, pero el conocimiento es distinto de la simple información e implica la contextualización e incorporación de otros elementos que doten a esta información de un sentido o utilidad (Ackoff, 1996; Boisot & Canals, 2004).

De acuerdo con las anteriores definiciones, resulta claro que para que un conjunto de datos pueda convertirse en conocimiento, es indispensable que éstos

puedan ser organizados, estructurados e incorporados dentro de un contexto específico.

Una forma natural mediante la cual el cerebro humano adquiere conocimiento es la **clasificación** de la información que vamos acumulando. De manera automática, el cerebro organiza la información que percibimos de acuerdo a ciertas características sobresalientes o importantes, integrándola con conocimientos adquiridos con anterioridad. De esta forma somos capaces de generar nuevo conocimiento de nuestras experiencias y percepciones tanto en el terreno meramente práctico, como en la actividad intelectual. Así bien, aunque existen muchos mecanismos y estrategias de organización y estructuración de la información, la clasificación es casi siempre un paso natural, básico y fundamental para convertir datos e información en conocimiento.

El proceso de clasificación, que nuestro cerebro realiza generalmente de manera automática, implica la extracción o selección de **características relevantes** de los datos y el agrupamiento de los mismos de acuerdo a dichas características. Una vez clasificados, los datos adquieren nueva información y se hace posible el descubrimiento de relaciones de mayor orden que, antes del proceso de clasificación, se encontraban ocultas. En la medida en que se hacen observaciones más profundas y detalladas y se va acumulando información, los nuevos datos deben no solamente organizarse y clasificarse, sino también incorporarse como parte del conocimiento ya adquirido.

1.1.2 Generación de Conocimiento en Biología

El avance de las ciencias ha involucrado siempre la acumulación de datos, que posteriormente se han podido interpretar y organizar de forma estructurada para ir formando una base de conocimiento sobre la que se han ido incorporando a su vez nuevos datos con sus interpretaciones y contextualizaciones. En particular la **Biología** es una ciencia donde la cantidad de datos generalmente ha superado la capacidad de los científicos para organizarlos y podemos decir que se ha desarrollado “a golpe de clasificaciones”. Es decir, algunos de los grandes avances en esta ciencia se han conseguido a partir de la clasificación de datos acumulados (Ouzounis CA et al., 2003).

El gran aporte de Aristóteles (considerado por muchos como padre de la biología) a las ciencias, fue la introducción del primer sistema de clasificación de los seres vivos. Así, la masa informe de plantas y animales se convirtió en un conjunto estructurado de organismos que compartían características dentro de un mismo grupo y se diferenciaban a su vez en otros rasgos de los organismos pertenecientes a otros grupos. Siglos después, surgiría la biología como una ciencia independiente gracias al trabajo de varios naturalistas cuyas aportaciones tienen su raíz también en el campo de la clasificación y me refiero aquí a los fundadores de la biología contemporánea: Linneo, Bonet, Lamarck y Darwin, por mencionar sólo a algunos de los más sobresalientes. La gran importancia del trabajo de estos científicos residió en su capacidad de extraer información relevante de los datos existentes y organizarlos de forma estructurada, generando así nuevo conocimiento. (O´Neil, 2005)

La tarea de clasificar, puede definirse como la actividad de agrupar los elementos de información o datos de acuerdo a propiedades o atributos comunes. Los datos son la materia prima de la información; son el resultado de investigación y descubrimiento. Un dato aislado no tiene significado a menos que se entienda su

contexto. Los datos deben transformarse en información para poder incorporarse como conocimiento útil.

Durante las últimas dos décadas, con el desarrollo de nuevas técnicas de biología molecular y los proyectos de secuenciación de genomas completos, ha habido una explosión en la cantidad de datos de secuencias de genes y proteínas. Esto nos ha inundado de información. Pero es indispensable encontrar la forma de organizarlos y clasificarlos para poder dilucidar relaciones de mayor orden y transformarlos en conocimiento (Bork P & Koonin E, 1998; Lewis SE, 2004). “El gran reto de la investigación en biología hoy es cómo convertir los datos en conocimiento” (Brenner S, 2002). Para deducir posibles pistas sobre la función e interacción de estas moléculas en las células, es necesario clasificarlas en categorías con significado que se liguen globalmente al conocimiento biológico existente.

1.2 El contexto biológico

1.2.1 Antecedentes

La **Biología Molecular** que, como su nombre lo indica, es el estudio de la vida a nivel molecular, surge a mediados del siglo XX de la interacción entre biofísica, genética y bioquímica, convergiendo en torno a un mismo problema: definir la estructura y función del gen. El objeto de estudio de la biología molecular lo constituyen las moléculas responsables de la transmisión y expresión de la información genética en las células, a saber: el ácido Desoxirribonucléico (**DNA**), el ácido Ribonucléico (**RNA**) y las proteínas, así como su estructura, función e interacciones.

Después de un periodo de intenso trabajo enfocado a comprender los mecanismos moleculares implicados en la transmisión y expresión de la información genética almacenada en las moléculas de DNA, comenzaron a desarrollarse técnicas que permitían conocer la secuencia de genes y proteínas, es decir, la identidad y orden de los elementos que los componen. Con el perfeccionamiento de la técnicas de secuenciación a finales de la década de los 1980´s, inició un periodo, que continúa hasta nuestros días, de generación masiva de datos a partir de los grandes proyectos de secuenciación de genomas completos (Lewis SE, 2004) [a la fecha (agosto de 2005) más de 160 genomas han sido totalmente secuenciados (Darden et al., 2005)]. El resultado de estos proyectos ha sido la generación de enormes volúmenes de datos que requieren ser procesados, organizados e interpretados para poder extraerles información adicional que permita comprender a profundidad las relaciones entre estructura y función y alcanzar una mayor comprensión sobre la claves de la evolución a nivel molecular. Es precisamente en el manejo de estas grandes bases de datos donde las ciencias de la computación se han incorporado a la biología molecular dentro del campo de la **bioinformática** conocido como **genómica computacional** (Kanehisa M & Bork P, 2003).

1.2.2 Conceptos básicos

La información genética almacenada en las moléculas de DNA de un organismo es conocida como su **genoma**. El DNA es una estructura lineal de doble hélice, compuesta por dos cadenas entrelazadas de bloques conocidos como **nucleótidos**. Cuatro nucleótidos constituyen el alfabeto básico de la molécula de DNA: Adenina (A), Timina (T), Guanina (G) y Citosina (C). Los **genes** son las regiones funcionales del DNA y deben ser leídos por la maquinaria celular para

generar su producto, que en la mayoría de los casos es una proteína[1]. El primer paso consiste en copiar o **transcribir** la información de la región de DNA correspondiente al gen en una molécula de RNA complementaria. La molécula de RNA resultante se conoce como RNA mensajero (**RNA_m**) y viaja al citoplasma, donde se lleva a cabo el proceso de la **traducción**, es decir, la generación de una proteína a partir de la plantilla definida en la molécula de RNA_m (fig.1).

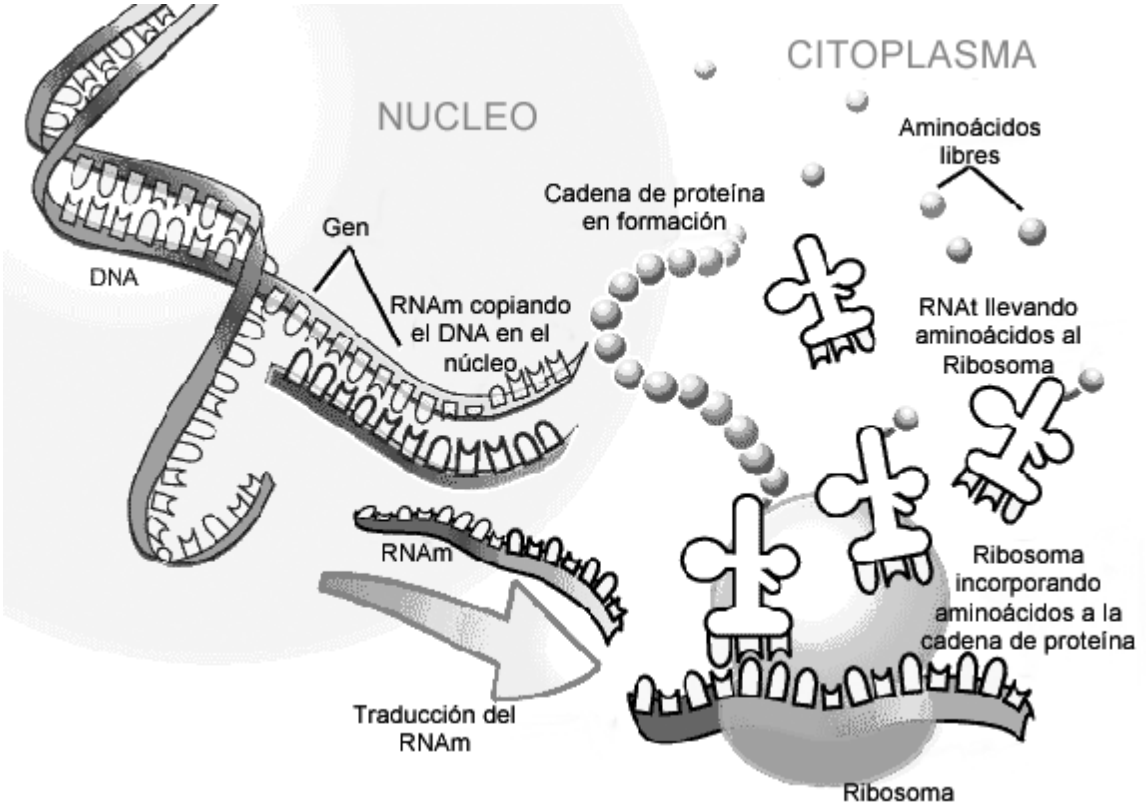


Fig. 1. Muestra los procesos de transcripción y traducción. (Imagen modificada de Wang, J http://www.bioteach.ubc.ca/MolecularBiology/AMonksFlourishingGarden/#Fig4).

Las proteínas son macromoléculas complejas que constituyen más del 50% de la composición celular de los organismos vivos. Podría decirse que así como los ácidos nucleicos son la base de la herencia, las proteínas son la base de la vida, ya que conforman la mayor parte de la maquinaria estructural y funcional de todas las células en los organismos vivos. La proteínas son responsables de controlar las condiciones fisicoquímicas al interior de las células, son los componentes básicos de la estructura celular, llevan a cabo el transporte y almacenamiento de pequeñas moléculas, están involucradas en la transmisión de señales biológicas y catalizan todas las reacciones bioquímicas de las células, en cuyo caso se denominan enzimas.(Griffiths et al., 2000)

[1] Existen también genes que codifican para RNA y regiones de DNA que aparentemente no tienen ningún producto génico asociado. En este trabajo utilizamos únicamente ORF´s (Open Reading Frames), es decir genes, o regiones de DNA que codifican para proteínas, con una Metionina inicial.

Las unidades estructurales de las proteínas son los **aminoácidos**. Aunque se conocen cerca de 180 aminoácidos que existen de forma natural en los organismos vivos, sólo 20 de éstos son utilizados para la construcción de proteínas (Lehninger A, 1982). Estos 20 aminoácidos son conocidos como aminoácidos frecuentes. Los demás aminoácidos se encuentran en diferentes células y tejidos en forma libre o combinada, pero nunca en las proteínas y generalmente actúan como precursores o intermediarios en el metabolismo. Aquí nos ocuparemos exclusivamente de estos 20 aminoácidos que constituyen el alfabeto básico de todas las proteínas conocidas.

La secuencia de aminoácidos de una proteína está determinada por la secuencia de nucleótidos del gen que la codifica. Cada aminoácido está codificado en la molécula de RNA_m por un grupo de tres nucleótidos, conocido como **codón**. El conjunto de reglas que especifica qué aminoácido está codificado por cada codón se conoce como **Código Genético** (fig. 2).

El Código Genético

	U		C		A		G		
U	UUU UUC	Fenil-alanina	UCU UCC UCA UCG	Serina	UAU UAC	Tirosina	UGU UGC	Cisteína	**U** **C**
	UUA UUG	Leucina			UAA UAG	ALTO	UGA UGG	ALTO Triptofano	**A** **G**
C	CUU CUC CUA CUG	Leucina	CCU CCC CCA CCG	Prolina	CAU CAC	Histidina	CGU CGC CGA CGG	Arginina	**U** **C**
					CAA CAG	Glutamina			**A** **G**
A	AUU AUC AUA	Isoleucina	ACU ACC ACA ACG	Treonina	AAU AAC	Aspara-gina	AGU AGC	Serina	**U** **C**
	AUG	Metionina			AAA AAG	Lisina	AGA AGG	Arginina	**A** **G**
G	GUU GUC GUA GUG	Valina	GCU GCC GCA GCG	Alanina	GAU GAC	Acido Aspártico	GGU GGC GGA GGG	Glicina	**U** **C**
					GAA GAG	Acido Glutámico			**A** **G**

Fig. 2. El Código Genético.

Los aminoácidos difieren en cuanto a sus propiedades fisicoquímicas (tamaño, polaridad, carga, hidrofobicidad, capacidad de generar cierto tipo de enlaces, etc.), y son estas características las que confieren estructura y función específica a cada proteína (fig. 3). Los aminoácidos se unen entre sí de manera secuencial mediante un tipo de enlace covalente conocido como enlace peptídico. La secuencia de aminoácidos de una proteína se conoce también como su **estructura primaria**. La longitud promedio de las proteínas es de 350 aminoácidos,

aunque las hay tan pequeñas como de una docena y tan largas como de miles (la proteína más larga que se conoce consta de 5000 aminoácidos).

De manera esquemática, cada aminoácido puede ser representado por una letra y las proteínas pueden ser vistas como palabras formadas por un alfabeto de 20 letras (fig. 3). De acuerdo con el dogma central del plegamiento de las proteínas, la estructura primaria de la proteína determina cómo se dará el plegamiento de la misma en tres dimensiones. Muchas fuerzas covalentes y no covalentes intervienen en la determinación de la estructura tridimensional de las proteínas, algunas de ellas son: interacciones de Van der Waals, fuerzas hidrofóbicas, fuerzas electrostáticas, momentos dipolares, puentes disulfuro y puentes de hidrógeno. Es la estructura tridimensional la que permite que la proteína realice su función biológica específica. Las reglas que gobiernan el plegamiento de las proteínas aún no se conocen en su totalidad.

AMINOACIDO	Cod 3letras	Cod 1letra
Alanina	Ala	A
Arginina	Arg	R
Asparagina	Asn	N
Aspartato	Asp	D
Cisteína	Cys	C
Fenilalanina	Phe	F
Glicina	Gly	G
Glutamato	Glu	E
Glutamina	Gln	Q
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lys	K
Metionina	Met	M
Prolina	Pro	P
Serina	Ser	S
Tirosina	Tyr	Y
Treonina	Thr	T
Triptofano	Trp	W
Valina	Val	V

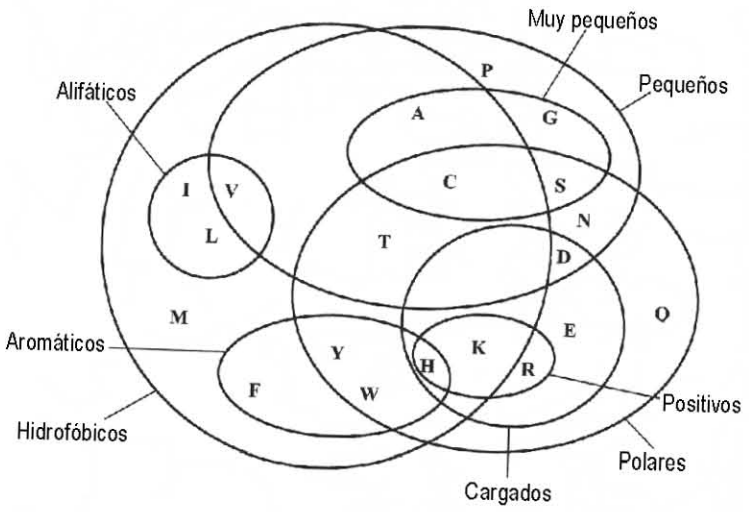


Fig. 3. Los 20 aminoácidos constitutivos de las proteínas, así como sus abreviaturas y propiedades fisicoquímicas.
(Imagen tomada de: http://www.hgmp.mrc.ac.uk/Courses/Intro_3day/Appendix5.html).

No todas las posibles combinaciones de aminoácidos pueden formar una estructura tridimensional estable y funcional. En el transcurso de la evolución se han seleccionado solo aquellas secuencias con una estructura estable. Una secuencia de aminoácidos siempre produce la misma estructura tridimensional, pero varias secuencias distintas pueden adoptar estructuras tridimensionales semejantes. La estructura tridimensional de las proteínas está más conservada en la evolución que la secuencia misma. Uno de los principales objetivos de la biología molecular es entender la estructura y función de las proteínas. La estructura de una proteína es lo

que proporciona mayor información sobre su función, pero determinar la estructura tridimensional de una proteína es un proceso complicado y no siempre posible. De todas las proteínas secuenciadas sólo se conoce la estructura para una pequeña fracción (Whisstock JC & Lesk AM, 2003). En la ausencia de datos estructurales, el análisis de la secuencia continúa siendo la principal fuente de información con la que contamos. (Yona G, 1999)

1.3 Análisis de Secuencias

1.3.1 Consideraciones generales del análisis de secuencias

Se han utilizado muchos criterios para poder clasificar las proteínas. Algunos de estos criterios se basan en propiedades físicas (solubilidad, hidrofobicidad, etc.), otros en propiedades químicas (polaridad, carga, enlaces) y otros más en forma y función. Todos estos métodos se desarrollaron en principio para clasificar proteínas de estructura y función conocidas. Sin embargo, después del "Big Bang" de la biología molecular en que cientos de miles de nuevas secuencias han sido descritas, se han tenido que desarrollar nuevas aproximaciones para tratar de organizar estas proteínas de las que se tiene poca o ninguna información más allá de su secuencia de aminoácidos.

Un concepto generalmente aceptado es que las proteínas que comparten la misma estructura primaria tienen también una similitud funcional (Guralnik V & Karypis G, 2001; Yona G et al., 1999). La detección de similitudes entre secuencias puede ayudar a revelar la función biológica de nuevas proteínas, así como su origen y relaciones con otras proteínas. La similitud de secuencia no es siempre fácil de detectar, ya que durante la evolución han cambiado por sustituciones, inserciones y deleciones. Algunos de estos eventos evolutivos pueden ser trazados mediante la utilización de algoritmos de comparación de secuencias. Cuando las proteínas tienen una alta similitud de secuencia generalmente se asume que tienen un ancestro común y se conocen como proteínas homólogas.

Dada una nueva secuencia de proteína, el enfoque básico para predecir su función recae en la comparación pareada de secuencias con otras proteínas cuyas propiedades ya sean conocidas (Althschul SF et al., 1990; Lipman DJ & Pearson WR, 1985a; Needleman SB & Wunsh CD, 1970; Smith TF & Waterman MS, 1981). Dichos métodos se han aplicado por décadas y han sido útiles para identificar la función biológica de muchas nuevas secuencias. Sin embargo, la cantidad de datos acumulados en los últimos años no puede seguirse analizando mediante comparaciones pareadas. Muchos nuevos métodos, la mayoría de los cuales se basan en técnicas de alineamiento múltiple, se han desarrollado para poder clasificar grandes volúmenes de información (Corpet F, 1988; Feng & Doolittle, 1987; Gribskov et al., 1987; Higgins DG et al., 1996; Higgins DG & Sharp PM, 1988; Johnson & Doolittle, 1986; Taylor WR, 1989).

1.3.2 Nuevos enfoques al análisis de secuencias

Los métodos que se han desarrollado para hacer análisis y clasificación de secuencias a gran escala, pueden agruparse de acuerdo a dos criterios. En términos del enfoque utilizado pueden dividirse en:

a) **Parciales**.- El análisis se centra en solo una parte de la secuencia,
b) **Globales**.- En los que se tratan de analizar las secuencias completas.

Y de acuerdo al nivel de automatización del proceso de clasificación, en:

a) **Supervisados**.- Se requiere la participación del experto para definir las clasificaciones,

b) **No supervisados**.- El proceso es automático y no requiere la intervención humana.

Los métodos parciales, parten de la hipótesis de que la función de la mayoría de las proteínas conocidas depende de pequeñas regiones estructurales o grupos de éstas, que son las responsables de que la proteína se pliegue y funcione de una manera específica. Estas regiones se habrían mantenido bastante conservadas en el transcurso de la evolución debido a la altísima presión de selección en contra de cualquier cambio en su estructura y por lo mismo, su identificación podría ser más fácil entre proteínas que hayan evolucionado, incluso a partir de un ancestro remoto. Dependiendo del tamaño y características de la región considerada para la clasificación, estos métodos se dividen en:

1) Métodos de "motifs", en los que la región considerada está constituida por pequeños fragmentos continuos de secuencias [PROSITE-motif (Bairoch, 1991; Bairoch, 1993), e-motif (Huang JY & Brutlag DL, 2001; Nevill-Manning CG et al., 1998)].

2) Métodos de dominios, en los que se consideran regiones funcionales, que pueden estar formadas por uno o varios motifs, no necesariamente contiguos en la secuencia, pero adyacentes en la estructura tridimensional y responsables de una función específica de la proteína [ProDom(Corpet F et al., 1998; Corpet F et al., 2000), DOMO (Gracy J & Argos P, 1998a; Gracy J & Argos P, 1998b)].

3) Métodos de múltiples motifs o firmas, que se basan en la localización de pequeñas regiones conservadas, separadas por espacios de longitud variable, no tan conservados [Blocks (Henikoff et al., 1999; Henikoff & Henikoff, 1991), Prints (Attwood et al., 1994; Attwood et al., 2003)].

4) Métodos de profiles, en los que se define una matriz de pesos, asociados a cada aminoácido en las distintas posiciones de la secuencia, para regiones identificadas como funcionalmente importantes en las proteínas [Pfam (Bateman et al., 2002; Sonnhammer ELL et al., 2005), SMART (Letunic I et al., 2002; Schultz J et al., 1998), PROSITE-profile (Sigrist CJA et al., 2002)].

La mayoría de estos métodos utilizan técnicas supervisadas o semi-supervisadas para establecer los límites entre grupos o familias de proteínas. Todos parten de algún tipo de alineamiento de las secuencias y cálculos de índices de similitud. Estos análisis han generado excelentes bases de datos de motifs y dominios que sirven para identificar patrones en nuevas secuencias, pero no permiten tener una visión global de las proteínas.

Los métodos globales intentan hacer una clasificación del universo de secuencias de proteínas tomándolas como una unidad. Son pocos los trabajos que aplican esta aproximación y en su mayoría utilizan algún tipo de Red Neuronal Artificial (Pasquier C et al., 2001), o técnicas de *clustering*, como k-means (Guralnik V & Karypis G, 2001) para clasificar las proteínas. En la mayoría de los casos parten de los resultados obtenidos de alineamientos múltiples (Enright AJ et al., 2003) y frecuentemente centran el análisis en pequeñas regiones estructurales más conservadas, es decir, aunque utilizan secuencias completas, el análisis se sesga hacia la presencia o ausencia de motifs comunes en las secuencias. Estos métodos

son en su mayoría supervisados. Entrenan las redes con proteínas previamente clasificadas por otros métodos y luego intentan generar una red capaz de clasificar de manera correcta nuevas secuencias de proteínas de función desconocida.

Sin embargo, el prerrequisito de tener grupos predefinidos como entrada, ocasiona que la mayoría de estos métodos no puedan ser aplicados a la totalidad de las secuencias de proteínas ya que hay muchas nuevas secuencias que no muestran similitud con ninguna de las ya estudiadas. Aún más, estos análisis no nos permiten tener una representación matemática de las secuencias de proteínas ni una visión global del espacio de secuencias. Dicha visión podría llevar al descubrimiento de características de orden superior en el espacio de secuencias de proteínas. Esto es extremadamente importante si consideramos que la mayoría de los métodos fallan al asignar función biológica a más del 40% de las secuencias (Yona G, 1999). Es en este sentido que consideramos que la utilización de técnicas de clasificación no supervisada podría aportar una nueva visión al problema de la clasificación de proteínas.

1.4 Métodos de clasificación no supervisada

1.4.1 Consideraciones generales

El ***clustering*** puede ser considerado como el problema de aprendizaje no supervisado más importante. El objetivo es encontrar una estructura en un conjunto de datos sin etiquetar, es decir, que no han sido previamente clasificados, ni existen ejemplos que determinen qué tipo de relaciones pueden ser válidas entre los datos (Halkidi M et al., 2001). Una definición laxa de *clustering* podría ser *"el proceso de organizar objetos en grupos cuyos miembros se asemejan en algún sentido"*. Un *cluster* sería entonces un conjunto de objetos que son semejantes entre sí y distintos de los objetos que pertenecen a otros conjuntos.

El *clustering* es la organización de una colección de patrones (usualmente representados como un vector de características o un punto en un espacio multidimensional), en grupos o clases, basados en su similitud (Jain AK et al., 1999). Es muy importante entender la diferencia entre *clustering* (clasificación no supervisada) y clasificación supervisada. En la clasificación supervisada se parte de una colección de patrones etiquetados (preclasificados) y el problema reside en poder etiquetar patrones nuevos de acuerdo con las clases ya definidas. Típicamente los patrones etiquetados son usados para aprender la descripción de las clases y esta información es usada para etiquetar los nuevos patrones. En el *clustering*, el problema consiste en agrupar un conjunto de patrones no etiquetados en grupos con significado. El agrupamiento de los patrones está dado exclusivamente por los datos mismos.

Existen muchas técnicas de clasificación no supervisada, pero se pueden establecer 4 pasos básicos que deben seguirse en todos los casos:

1) Representación de patrones, que implica la selección y extracción de características.

2) Definición de una medida de proximidad (similitud), apropiada a los datos.

3) *Clustering* (o agrupamiento), que implica la elección de un algoritmo que nos proporcione un esquema de clasificación apropiado para nuestros datos.

4) Validación, que implica la verificación de la validéz de los *clusters*, usando técnicas y criterios apropiados.

La diferencia más importante reside en el método seleccionado para llevar a cabo el *clustering* o agrupamiento de los datos. En términos muy generales, podemos decir que el *clustering* puede clasificarse en: a) **Exclusivo** y b) **Difuso**.

Estas dos formas de clasificación dependen de si los objetos clasificados pertenecerán de manera exclusiva a una de las clases o si pueden tener grados de pertenencia a las distintas clases.

Otra taxonomía divide los métodos en: a) **Jerárquico** y b) **Partitivo**, dependiendo de si los *clusters* resultantes se agrupan entre sí para formar una estructura tipo árbol o se mantienen independientes unos de otros. El método más apropiado dependerá de la naturaleza de los datos y el objetivo que se persiga.

1.4.2 Mapas Autoorganizados (SOMs)

Dentro de los métodos de clasificación no supervisados, partitivos y exclusivos, se encuentran los **mapas autoorganizados** o **redes de Kohonen** (Kohonen T, 1990). El método no requiere hacer ninguna suposición *a priori* respecto a los datos que vamos a clasificar y su objetivo es descubrir la estructura subyacente a los mismos.

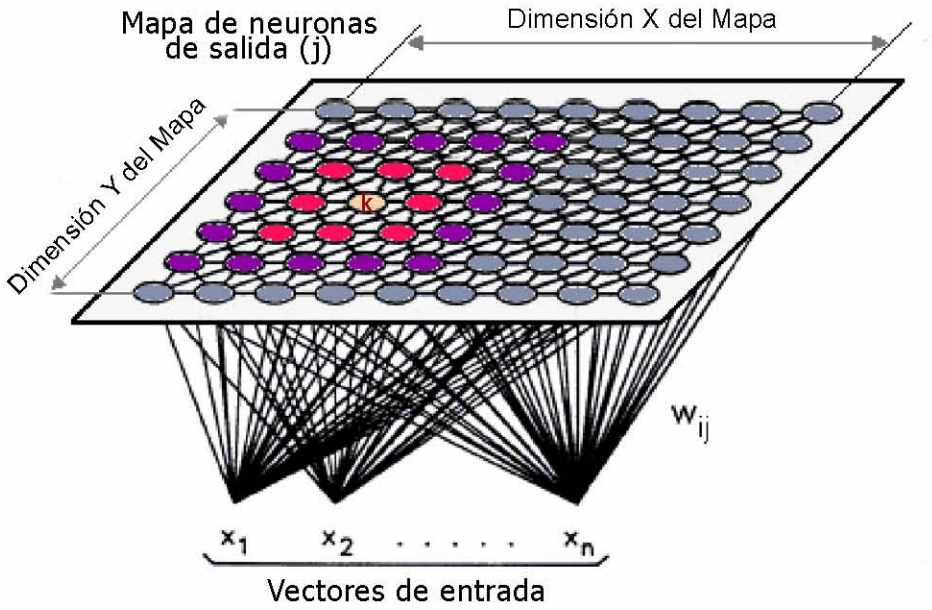


Fig. 4. Estructura básica de una red de Kohonen. Las entradas representadas por un vector multidimensional y las salidas dispuestas como un grid bidimensional.
Se muestra también la zona de vecindad de la neurona ganadora (en amarillo).
(Imagen tomada de: http://www.lohninger.com/helpcsuite/img/kohonen1.gif)

Los mapas autoorganizados (SOM) son redes neuronales artificiales (ANN) competitivas. Poseen una arquitectura de dos capas (entrada-salida), funciones de activación lineales y flujo de información unidireccional. Las unidades de entrada reciben datos continuos normalizados y también se normalizan los pesos de las conexiones con las capas de salida. Tras el aprendizaje de la red, cada patrón de entrada activará una única unidad de salida (*clustering* exclusivo). Las neuronas de salida pueden estar ordenadas en forma de un arreglo bidimensional, de forma que el resultado del proceso de clasificación podría también verse como el mapeo en

dos dimensiones de un conjunto de patrones multidimensionales (fig.4). Estos mapas son el resultado de una compresión de la información, ya que retienen solamente las características comunes más relevantes del conjunto de señales de entrada (Ferran E et al., 1994)

El objetivo de este tipo de redes es clasificar los patrones de entrada en grupos de características similares, de modo que cada grupo activará siempre las mismas salidas. En cada ciclo de aprendizaje, las neuronas de la capa de salida compiten por activarse y solo una de ellas permanece activa ante una determinada información de entrada a la red. Cuando una neurona resulta ganadora, no solo su peso es ajustado, sino también el de todas las neuronas vecinas, creándose así zonas de influencia de las neuronas vencedoras. De esta forma, la configuración final del mapa bidimensional es importante porque refleja relaciones de semejanza entre las clases vecinas.

1.4.2.1 Algoritmo de entrenamiento

Aunque existen muchas variantes del algoritmo de Kohonen, aquí describiremos el que aplica el programa Data Engine, que se utilizó en este trabajo (MIT GmbH, 1997).

1) Se inicializan los pesos entre las neuronas (**i**) de la capa de entrada y las neuronas (**j**) de la capa de salida $\mathbf{w_{ij}}$ con valores aleatorios.

2) Se selecciona al azar un vector de entrada $\mathbf{x_i}$.

3) Se encuentra la neurona ganadora **k**, que será aquella cuya distancia euclidiana con respecto al vector de entrada $\mathbf{x_i}$ sea mínima (**d** = min $||\mathbf{x_i}-\mathbf{w_{ij}}||$).

4) Se ajustan los vectores de peso de acuerdo con la siguiente función:

$$\mathbf{w_{ij}'} = \mathbf{w_{ij}} + \boldsymbol{\alpha}\mathbf{(t)} \cdot \mathbf{r_{jk}(t)} \cdot (\mathbf{x_i} - \mathbf{w_{ij}})$$

5) Se regresa al paso (2).

Donde: **α(t)** es la tasa de aprendizaje en la época **(t)** y

$\mathbf{r_{jk}(t)}$ es la función de propagación a la neurona **j** desde la neurona ganadora **k** en la época **(t).**

Los factores de propagación que determinan la dependencia espacial de las neuronas en el mapa de Kohonen se definen utilizando la función Gaussiana: $r_{jk} = e^{\left(-\frac{d^2}{\sigma^2}\right)}$

Donde: **σ**(t) es el radio de aprendizaje en la época (t) y

d es la distancia (|| **j**-**k** ||) entre la neurona **j** y la neurona ganadora **k** en el mapa de Kohonen. La distancia hacia las vecinas y perpendiculares, canónicamente, se hace igual a 1.0. De allí se desprende que la distancia entre vecinas en diagonal es de longitud $\sqrt{2}$ y así sucesivamente.

Tanto la tasa de aprendizaje **α** como el radio de aprendizaje **σ** dependen también de la época (t). Después de cada ciclo de entrenamiento se recalculan, multiplicándolas por un factor f_{α} y f_{σ} como sigue: $\alpha(t) = f_{\alpha} \cdot \alpha(t-1)$ y $\sigma(t) = f_{\sigma} \cdot \sigma(t-1)$. De esta definición operativa se desprende que ambas decrecen exponencialmente durante el entrenamiento.

2

OBJETIVO

Un reto de la Biología Molecular es poder descubrir las reglas que gobiernan la expresión y transmisión de la información genética en los organismos vivos. Si bien se conocen ya muchos de los mecanismos implicados en el proceso, una de las grandes asignaturas pendientes es dilucidar la relación que existe entre la secuencia de aminoácidos y la estructura y función de las proteínas.

Actualmente se cuenta con una enorme cantidad de datos de secuencias de proteínas, pero para muchas de éstas no se cuenta con ninguna información adicional, no se sabe dónde y cuando se expresan, qué estructura tridimensional adoptan ni que función desempeñan en el organismo. Una forma de aproximarse a este problema es precisamente la clasificación de los datos, organizarlos y estructurarlos para poder transformarlos en conocimiento. Con los métodos de análisis aplicados tradicionalmente a este problema se obtiene solo una visión parcial. Es necesario encontrar nuevas formas de enfocar el problema para poder obtener una visión más completa del universo de las proteínas.

El principal objetivo de este trabajo es explorar un nuevo enfoque al problema de la clasificación de proteínas que pueda ayudar a encontrar posibles organizaciones de alto nivel en el espacio de secuencias e identificar conjuntos de proteínas relacionadas. Mediante la utilización de métodos de clasificación no supervisados se pretende obtener una clasificación no sesgada, es decir, permitir que las proteínas se agrupen entre sí en términos de características propias de las secuencias mismas, sin establecer limitaciones o criterios *a priori* que puedan dirigir la clasificación en algún sentido. En primera instancia, se pretende conseguir una clasificación de todas las secuencias en grupos relacionados, que podrían permitirnos asignar estructura y función a proteínas de función desconocida, pero más allá de esto, el método podría utilizarse también como punto de partida para la identificación de patrones o motivos característicos de cada clase, que pudieran proporcionar alguna pista sobre las claves ocultas en la determinación de la estructura y función de las proteínas (Kuri-Morales AF & Ortíz-Posadas MR, 2006), aportando un nuevo enfoque al campo de la proteómica.

Para alcanzar este objetivo proponemos un método novedoso para la generación de los vectores de características, que integra información de la composición y estructura de las secuencias, y la utilización de mapas autoorganizados para la generación de las clases, que esperamos revelen relaciones de orden superior entre las secuencias clasificadas.

Este trabajo se enmarca dentro del campo de la Biología Molecular Computacional o Bioinformática y tiene raíces en dos diferentes disciplinas: la Ciencia de la Computación y la Biología Molecular.

De manera muy concisa, podríamos resumir los objetivos en tres puntos:

1) Obtener una representación vectorial de la información contenida en los datos de secuencia de las proteínas de *S*. *cerevisiae*.

2) Clasificar las secuencias de proteínas utilizando métodos no sesgados.

3) Validar las clases encontradas mediante criterios funcionales.

3

METODOLOGÍA

En esta sección detallaré el procedimiento utilizado para alcanzar los objetivos anteriormente descritos. En la primera sección (3.1) explicaré de dónde obtuvimos los datos de secuencias con los que se trabajó. A continuación (3.2) describiré el método que utilizamos para la representación de las secuencias de proteínas en forma de vectores de características, que puedan ser utilizados para alimentar una red de Kohonen. Posteriormente (3.3) describiré el método utilizado para definir el número óptimo de grupos o *clusters*, así como el algoritmo utilizado para la clasificación. Por último (3.4), describiré la herramienta que diseñamos para la validación de nuestros grupos y las pruebas estadísticas que se utilizaron.

3.1 Obtención de las secuencias

Para explorar las potencialidades del método buscamos trabajar con un conjunto no sesgado de secuencias de proteínas. Con este fin decidimos utilizar el conjunto completo de los genes que codifican para proteínas[2] (ORF´s) de una especie cuyo genoma estuviera totalmente secuenciado.

Se escogió la levadura *Saccharomyces cerevisiae* por ser uno de los organismos mejor conocidos. Fue el primer eucarionte que se secuenció totalmente (1996). Se conocen 6700 secuencias de proteínas (Open Reading Frames ORF´s) y comparte aproximadamente 23% de su genoma con el humano. Las levaduras son hongos unicelulares que pertenecen a la división Ascomycota y al Phyllum Eukaryota.

Más allá de su valor industrial, S. *cerevisiae* es un organismo modelo muy útil para los científicos. Es eucarionte (su material genético está contenido en un núcleo), su ciclo celular es muy semejante al de las células de organismos superiores y está regulado por proteínas homólogas. Debido a que ha sido utilizado por décadas como modelo experimental en todo tipo de estudios de bioquímica, biología celular y biología molecular, se ha acumulado una gran cantidad de información genética y bioquímica, que se pensó podría ser de utilidad en la fase de validación de los resultados.

La mejor fuente de información genética y de biología molecular de levadura es el sitio Web: “Saccharomyces Genome Database” (http://www.yeastgenome.org), mantenido por la Escuela de Medicina de la Universidad de Stanford. De ahí obtuvimos el archivo en formato FASTA (Lipman DJ & Pearson WR, 1985b) para todos los ORF´s conocidos de la especie. En la Fig.5 se muestra la secuencia de uno de los genes utilizados en formato FASTA, donde la primera línea corresponde a una etiqueta con identificadores del gen y a partir de la segunda línea se presenta la secuencia de aminoácidos de la proteína utilizando el código de representación de una sola letra por aminoácido.

[2] A lo largo de este trabajo utilizaremos indistintamente la palabra gen y proteína, ya que el archivo que utilizamos para hacer la clasificación es exclusivamente de genes que codifican para proteínas, también conocidos como ORF´s.

```
>YAL009W SPO7 SGDID:S0000007
MEPESIGDVGNHAQDDSASIVSGPRRRSTSKTSSAKNIRNSSNISPASMIFRNLLILEDD
LRRQAHEQKILKWQFTLFLASMAGVGAFTFYELYFTSDYVKGLHRVILQFTLSFISITVV
LFHISGQYRRTIVIPRRFFTSTNKGIRQFNVKLVKVQSTWDEKYTDSVRFVSRTIAYCNI
YCLKKFLWLKDDNAIVKFWKSVTIQSQPRIGAVDVKLVLNPRAFSAEIREGWEIYRDEFW
AREGARRRKQAHELRPKSE
```

Fig. 5. Secuencia de aminoácidos en formato FASTA para el gen YAL009W de *Saccharomyces cerevisiae*.

3.2 Representación de las secuencias

3.2.1 Planteamiento general

El primer paso para poder aplicar cualquier método de clasificación no supervisada consiste en la selección de características de nuestros datos y su representación en forma de un vector multidimensional. El objetivo es seleccionar de manera adecuada las características sobre las que se llevará a cabo la clasificación, de forma que se codifique la mayor cantidad de información posible y relevante de los objetos a clasificar (Halkidi M et al., 2001).

La secuencia de aminoácidos de una proteína se conoce como su **estructura primaria**. La **estructura secundaria** esta caracterizada por un conjunto finito de elementos estructurales formados por pequeños segmentos de la cadena de aminoácidos (30 a 40 residuos), que adoptan conformaciones características, tales como hélices alfa u hojas beta, debido al establecimiento de enlaces entre los distintos aminoácidos de la cadena. Las estructuras secundarias se empacan y establecen nuevos plegamientos para conformar lo que se conoce como la **estructura terciaria** de la proteína. De acuerdo al dogma central del plegamiento de las proteínas, la secuencia de aminoácidos determina la forma en que la proteína se plegará en tres dimensiones y es la estructura tridimensional la que permite que la proteína desempeñe su función biológica específica. Es razonable asumir que debe existir un mapeo entre la estructura primaria, secundaria y terciaria y la función biológica de la proteína. En lugar de tratar de encontrar este mapeo de manera directa, proponemos establecer un mapa de entidades de una, dos y tres dimensiones a un plano matemático (que llamaremos **Λ**) y luego mapear dichas entidades a un plano biológico (que llamaremos **Θ**).

Al mapeo de la secuencia de aminoácidos al plano **Λ** lo llamaremos “mapeo **λ**” y al mapeo de **Λ** a **Θ** lo llamaremos “mapeo **π**”. Así el razonamiento que subyace a nuestro método es encontrar entidades matemáticas de una, dos y tres dimensiones, de forma que la analogía entre dichas entidades y la estructura primaria, secundaria y terciaria de las proteínas pueda emerger.

3.2.2 Mapeo λ

Para explicar la forma en que construimos el vector de características, utilizaremos como ejemplo la secuencia de un gen cualquiera de nuestro archivo. En particular, el gen YAL009W que presentamos en la fig.5, nos permite describir el método de forma sencilla, por ser de pequeña longitud. Sea L (L=259) la longitud de la cadena de aminoácidos.

El vector de características para cada secuencia de proteína estará representado por 20 meta-columnas, cada una correspondiente a uno de los 20 aminoácidos conocidos (fig.3). La información para cada aminoácido se obtiene mediante la representación consecutiva de la cadena en la forma de arreglos de

una, dos, tres y cuatro dimensiones. El primer elemento de cada meta-columna siempre corresponde a la frecuencia relativa del aminoácido en la cadena (fig.6). Así, si consideramos el caso del aminoácido **H** (Histidina) en la cadena del ejemplo (fig.5), vemos que aparece en 5 ocasiones (**n**=5) en una cadena de longitud 259 (**L**=259), por lo tanto su frecuencia relativa será de (**n**/**L**) 5/259= 0.0193. Este parámetro se normaliza asignando una frecuencia de 1 al aminoácido más frecuente en todas las cadenas analizadas. Los demás elementos del vector corresponderán a la posición promedio y la desviación media de la posición en cada dimensión, incorporándose la información correspondiente a cada uno de los arreglos de acuerdo a los siguientes pasos:

A (Alanina)	C (Cisteína)	D (Aspartato)	E (Glutamato)	F (Fenilalanina)	...(14)...	Y (Tirosina)
$f_{A...}$	$f_{C...}$	$f_{D...}$	$f_{E...}$	$f_{F...}$	...	$f_{Y...}$

Fig. 6. Primer paso de la construcción del vector de características. El primer elemento de cada metacolumna corresponde a la frecuencia relativa (f) del aminoácido en la cadena.

3.2.2.1 Una dimensión

a) Posición promedio (X1) - Para cada aminoácido se cuentan el número de ocurrencias en la cadena (**n**) y la posición de cada ocurrencia medida con respecto al inicio de la cadena ($\mathbf{x_i}$). Luego se calcula la posición promedio del aminoácido de acuerdo con la siguiente fórmula: $\bar{x} = \frac{1}{n}\sum^{n} x_i$. Y por último, se divide entre la longitud total de la cadena (L), para tener una posición relativa, independiente de la longitud de la secuencia.

Para el caso de la histidina (**H**), del ejemplo que estamos siguiendo, las posiciones en las que aparece son: 12, 66, 104, 123 y 252. De aquí que:

$\bar{x}_H = \frac{1}{5}(12+66+104+123+252) = \frac{1}{5}557 = 111.4$. Dividido entre la longitud de la cadena (259), tenemos que $X1_H = 111.4/259 = 0.4301$.

b) Desviación media de la posición ($\bar{\sigma}_{X1}$)– Para cada aminoácido de la cadena, se calcula también la desviación media respecto a la posición promedio de acuerdo a la siguiente expresión: $\bar{\sigma}_{\bar{x}} = \frac{1}{n}\sum^{n} |x_i - \bar{x}|$. También en este caso, normalizamos la desviación dividiendo entre la longitud total de la cadena.

Siguiendo con el ejemplo de **H** que hemos analizado. Haciendo las sustituciones correspondientes en la ecuación, tendríamos la siguiente expresión:

$$\bar{\sigma}_{\bar{x}H} = \frac{1}{5}\left(|12-111.4| + |66-111.4| + |104-111.4| + |123-111.4| + |252-111.4|\right) =$$

$$= \frac{1}{5}(99.4 + 45.4 + 7.4 + 11.6 + 140.6) = \frac{304.4}{5} = 60.88$$

Dividido entre la longitud de la cadena (259), tenemos que $\bar{\sigma}_{X1H}$ =60.88/259=0.235.

La fig. 7 muestra como se iría conformando el vector de características al incorporar la información para una dimensión.

A (Alanina)	C (Cisteína)	D (Aspartato)	E (Glutamato)	F (Fenilalanina)	...	Y (Tirosina)
f_A,X1,$\sigma_{X1...}$	f_C,X1,σ_{X1}...	f_D,X1,$\sigma_{X1...}$	f_E,X1,$\sigma_{X1...}$	f_F,X1,$\sigma_{X1...}$	...	f_Y,X1,$\sigma_{X1...}$

Fig. 7. Vector de características después de incorporarle la información correspondiente a una dimensión.

3.2.2.2 Dos dimensiones

Se procede a ordenar la cadena en un plano bidimensional[3] de NxN (donde N=$\lceil\sqrt{L}\rceil$), como se muestra en la fig.8. Colocando el primer aminoácido de la cadena en las coordenadas (1,1) del plano y procediendo de manera consecutiva con la secuencia hasta la posición (N,1). El aminoácido N+1 se coloca a continuación en la posición (1,2) y así sucesivamente hasta llegar a la posición (N,2). El 2N+1 se coloca en las coordenadas (1,3) y se sigue ordenando la cadena de esta forma hasta haber colocado el último aminoácido en el plano.

a) Posición promedio (X2, Y2) – Para el caso de dos dimensiones la posición promedio del aminoácido estará compuesta de dos elementos: la posición promedio en X y la posición promedio en Y. Calculamos cada una de manera independiente.

Si definimos x_i como la posición en el eje X de cada una de las (n) apariciones del aminoácido en cuestión, tenemos que: $\bar{x} = \frac{1}{n}\sum^{n} x_i$. Luego dividimos entre la dimensión del arreglo en X (N), para obtener un número entre 0 y 1: $X2 = \frac{\bar{x}}{N}$.

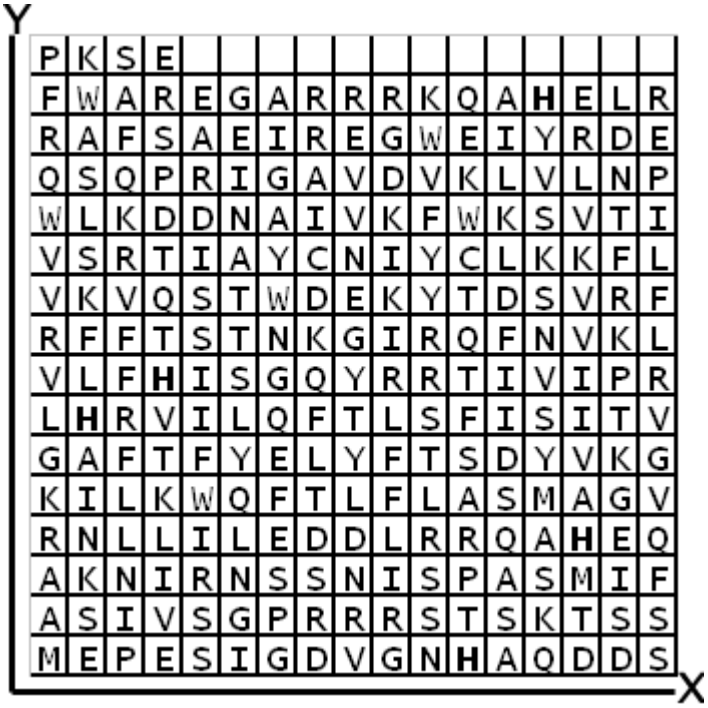


Fig. 8. Secuencia de aminoácidos del gen YAL0009W dispuesta en forma de un arreglo bidimensional.

[3] Escogimos la función $\sqrt{}$ para establecer las dimensiones del arreglo bidimensional porque nos pareció la más inmediata, pero también pudo haberse utilizado otra función.

Siguiendo con el ejemplo del aminoácido H en la cadena de la proteína YAL009W, podemos calcular X2, sustituyendo los valores correspondientes en la expresión anterior:

$\bar{x}_H = \frac{1}{5}(12+15+2+4+14) = \frac{1}{5}(47)) = \frac{47}{5} = 9.4$. Dividiendo entre la dimensión del arreglo en X (17), tenemos: $X2_H = \frac{9.4}{17} = 0.5528$

Para el caso de Y2, definimos x_j como la posición en el eje Y de cada una de las apariciones (n) del aminoácido y la posición promedio en Y será: $\bar{y} = \frac{1}{n}\sum^{n} x_j$. Dividiendo entre la dimensión del arreglo en Y (N), tenemos la siguiente expresión: $Y2 = \frac{\bar{y}}{N}$.

Nuevamente, podemos calcular Y2 para el aminoácido H de la cadena del ejemplo:

$\bar{y}_H = \frac{1}{5}(1+4+7+8+15) = \frac{35}{5} = 7$. Entre la dimensión del arreglo en Y (16), tenemos: $Y2_H = \frac{7}{16} = 0.4375$.

b) <u>Desviación media de la posición</u> ($\bar{\sigma}_{X2}$, $\bar{\sigma}_{Y2}$) – Como en el caso de una dimensión, calculamos la desviación como el promedio de las desviaciones absolutas con respecto a la posición promedio en cada uno de los ejes.

La expresión para X sería: $\bar{\sigma}_{\bar{x}} = \frac{1}{n}\sum^{n} |x_i - \bar{x}|$. Y nuevamente normalizamos dividiendo entre la dimensión del arreglo en X: $\bar{\sigma}_{X2} = \frac{\bar{\sigma}_{\bar{x}}}{N}$. La desviación de la posición en el eje Y, sería una expresión análoga a la anterior.

Siguiendo con el ejemplo del aminoácido H, tenemos que la desviación media de la posición en X sería:

$\bar{\sigma}_{\bar{x}H} = \frac{1}{5}(|12-9.4|+|15-9.4|+|2-9.4|+|4-9.4|+|14-9.4|) = \frac{25.6}{5} = 5.12$ y normalizando con respecto a la dimensión en X: $\bar{\sigma}_{X2H} = \frac{5.12}{17} = 0.3012$.

Para calcular la desviación de la posición en Y, sustituimos por los valores correspondientes:

$\bar{\sigma}_{\bar{y}H} = \frac{1}{5}(|1-7|+|4-7|+|7-7|+|8-7|+|15-7|) = \frac{18}{5} = 3.6$ y normalizando con respecto a la dimensión en Y (16): $\bar{\sigma}_{Y2H} = \frac{3.6}{16} = 0.225$.

La fig.9 muestra el vector de características con los elementos correspondientes a una y dos dimensiones

A	C	D	...	Y
$f_{A,}X1,\sigma_{X1},X2,\sigma_{X2},Y2,\sigma_{Y2...}$	$f_{C,}X1,\sigma_{X1},X2,\sigma_{X2},Y2,\sigma_{Y2...}$	$f_{D,}X1,\sigma_{X1},X2,\sigma_{X2},Y2,\sigma_{Y2...}$	...	$f_{Y,}X1,\sigma_{X1},X2,\sigma_{X2},Y2,\sigma_{Y2...}$

Fig. 9. Plantilla del vector de características después de incorporar la información para dos dimensiones.

3.2.2.3 Tres dimensiones

Para incorporar la información de la tercera dimensión, el primer paso consiste en ordenar la cadena de aminoácidos en un espacio tridimensional[4] de dimensiones NxNxN, donde N=$\left\lceil \sqrt[3]{L} \right\rceil$. Con esta finalidad, los aminoácidos se van ordenando de manera secuencial hasta llenar cada plano de dos dimensiones y así sucesivamente hasta completar el espacio tridimensional, como se muestra en la fig.10.

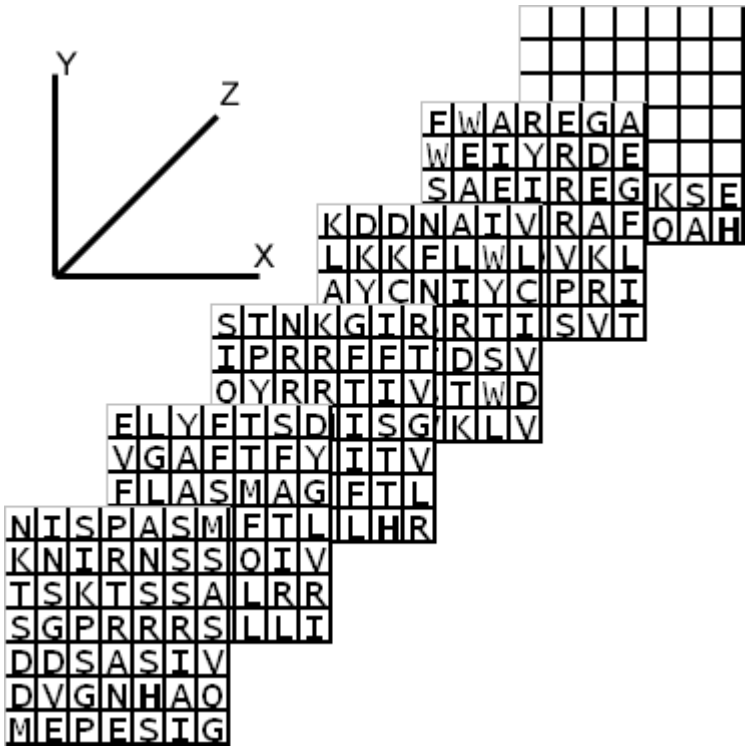


Fig. 10. Disposición de la secuencia del gen YAL009W en forma de un espacio tridimensional.

[4] Se utilizó la función $\sqrt[3]{\ }$ para establecer las dimensiones del arreglo tridimensional, por consistencia con la función elegida para el caso de dos dimensiones y con el mismo criterio utilizamos la raíz cuarta de la longitud para establecer las dimensiones del arreglo en cuatro dimensiones.

a) Posición promedio (X3, Y3, Z3) – Aquí, la posición promedio de cada aminoácido está conformada por tres componentes, que corresponderían a su posición en cada uno de los ejes (X,Y,Z) de espacio tridimensional en que hemos acomodado la cadena de aminoácidos. El cálculo de cada una de estas componentes del vector de características se hace de manera análoga a como lo hicimos para dos dimensiones.

Como ejemplo, pondremos la ecuación para calcular la posición promedio en Z. Sea z_i la posición en el eje Z de cada una de las ocurrencias de un aminoácido y n el número de ocurrencias del mismo: $\bar{z} = \frac{1}{n}\sum^{n} z_i$. Y normalizando: $Z3 = \frac{\bar{z}}{N}$, donde N es la dimensión en el eje Z del arreglo tridimensional de la secuencia.

Siguiendo con el ejemplo del aminoácido H en la secuencia modelo, calculamos los componentes X3, Y3 y Z3 del vector de características de la siguiente forma:

$\bar{x}_H = \frac{1}{5}(5+3+6+4+7) = \frac{25}{5} = 5$ y normalizando: $X3_H = \frac{5}{7} = 0.7143$

$\bar{y}_H = \frac{1}{5}(2+3+1+4+1) = \frac{11}{5} = 2.2$ y normalizando: $Y3_H = \frac{2.2}{7} = 0.3143$

$\bar{z}_H = \frac{1}{5}(1+2+3+3+6) = \frac{15}{5} = 3$ y normalizando: $Z3_H = \frac{3}{6} = 0.5$.

b) Desviación media de la posición ($\bar{\sigma}_{X3}$, $\bar{\sigma}_{Y3}$, $\bar{\sigma}_{Z3}$) – De la misma forma en que lo hemos definido para una y dos dimensiones, la desviación media será el promedio de las desviaciones absolutas con respecto a la posición promedio para cada una de las tres dimensiones (X, Y, Z).

La desviación media para H de nuestro ejemplo en las tres dimensiones será:

$\bar{\sigma}_{\bar{x}H} = \frac{1}{5}(|5-5|+|3-5|+|6-5|+|4-5|+|7-5|) = \frac{6}{5} = 1.2$

normalizando: $\bar{\sigma}_{X3H} = \frac{1.2}{7} = 0.17$

$\bar{\sigma}_{\bar{y}H} = \frac{1}{5}(|2-2.2|+|3-2.2|+|1-2.2|+|4-2.2|+|1-2.2|) = \frac{5.2}{5} = 1.04$ y

normalizando: $\bar{\sigma}_{Y3H} = \frac{1.04}{7} = 0.1486$

$\bar{\sigma}_{\bar{z}H} = \frac{1}{5}(|1-3|+|2-3|+|3-3|+|3-3|+|6-3|) = \frac{6}{5} = 1.2$ y normalizando:

$\bar{\sigma}_{Z3H} = \frac{1.2}{6} = 0.2$.

Con este mismo procedimiento se pueden ir incorporando dimensiones adicionales al vector de características. En la fig.11 se muestra un fragmento del vector de características, incorporando una, dos y tres dimensiones.

A	...	Y
f_A,X1,σ_{X1},X2,σ_{X2},Y2,σ_{Y2},X3,σ_{X3},Y3,σ_{Y3},Z3,σ_{Z3}	...	f_Y,X1,σ_{X1},X2,σ_{X2},Y2,σ_{Y2},X3,σ_{X3},Y3,σ_{Y3},Z3,σ_{Z3}

Fig. 11. Fragmento de la plantilla del vector de características incorporando una, dos y tres dimensiones.

De acuerdo con el método descrito en los párrafos anteriores, elaboramos un programa en perl (*mapeo_lambda.pl*) que a partir de un archivo de entrada con las secuencias de aminoácidos de las 6700 proteínas de S. *cerevisiae*, construye matrices de vectores de características con información de una dimensión Ξ_1, una y dos dimensiones Ξ_2, una, dos y tres dimensiones Ξ_3 y hasta cuatro dimensiones acumuladas Ξ_4. De esta forma conformamos matrices de 6700 renglones por 60 columnas, para el caso de Ξ_1 (20 AA x 3 características (f, X1, $\overline{\sigma}_{X1}$)), 140 columnas, para el caso de Ξ_2 (20 AA x 7 características (f, X1, $\overline{\sigma}_{X1}$, X2, $\overline{\sigma}_{X2}$, Y2, $\overline{\sigma}_{Y2}$)), 260 para Ξ_3 y 420 para Ξ_4. En la fig.12 se muestra un fragmento de la matriz Ξ_3.

	f_A	X1	σ_{X1}	X2	σ_{X2}	Y2	σ_{Y2}	X3	σ_{X3}	Y3	σ_{Y3}	Z3	σ_{Z3}	f_C	X1	σ_{X1}	X2
YEL054C	0.889	0.425	0.417	0.519	0.502	0.452	0.427	0.688	0.503	0.625	0.556	0.475	0.438	0.056	0.855	0	0.846
YEL008C-A	0.667	0.433	0.134	0.667	0.352	0.5	0.21	0.75	0	0.667	0.741	0.5	0.333	0.667	0.55	0.302	0.75
YER137C	0.364	0.62	0.491	0.558	0.487	0.625	0.503	0.792	0.394	0.6	0.444	0.7	0.55	0.182	0.664	0.655	0.558
YOR166C	0.419	0.477	0.455	0.593	0.559	0.492	0.467	0.688	0.525	0.507	0.573	0.493	0.406	0.163	0.478	0.357	0.675
YOR073W-A	0.063	0.364	0	0.111	0	0.444	0	0.6	0	0.5	0	0.5	0	0.063	0.727	0	0.222
YHL006C	0.318	0.519	0.355	0.56	0.471	0.536	0.364	0.548	0.438	0.667	0.529	0.514	0.31	0.045	0.793	0	0.154
YFL011W	0.604	0.528	0.438	0.544	0.571	0.542	0.451	0.503	0.51	0.504	0.601	0.57	0.411	0.226	0.518	0.482	0.542
YMR237W	0.519	0.506	0.48	0.486	0.605	0.521	0.498	0.553	0.583	0.593	0.506	0.553	0.472	0.123	0.626	0.436	0.496
YDL097C	0.526	0.467	0.455	0.457	0.48	0.486	0.476	0.55	0.544	0.525	0.544	0.529	0.446	0.105	0.583	0.315	0.714
YLR200W	0.25	0.537	0.674	0.764	0.215	0.527	0.657	0.44	0.462	0.56	0.427	0.6	0.64	0.05	0.719	0	0.455
YMR294W-A	0.462	0.405	0.336	0.545	0.576	0.439	0.361	0.633	0.667	0.667	0.741	0.467	0.267	0	0	0	0
YGR066C	0.355	0.549	0.61	0.545	0.39	0.551	0.625	0.545	0.64	0.61	0.614	0.621	0.601	0.194	0.669	0.478	0.685
YGL049C	0.539	0.508	0.513	0.517	0.461	0.515	0.527	0.569	0.51	0.56	0.677	0.513	0.457	0.056	0.733	0.089	0.426
YKL160W	0.143	0.738	0.287	0.564	0.469	0.722	0.311	0.833	0.247	0.6	0	0.8	0.267	0.238	0.262	0.139	0.723
YDR434W	0.379	0.53	0.532	0.55	0.498	0.532	0.536	0.64	0.624	0.575	0.671	0.55	0.512	0.061	0.416	0.487	0.51
YOL083W	0.275	0.49	0.597	0.424	0.557	0.509	0.59	0.489	0.441	0.489	0.386	0.532	0.501	0.175	0.391	0.409	0.667
YPL041C	0.143	0.671	0.315	0.467	0.507	0.7	0.336	0.767	0.533	0.7	0.533	0.7	0.347	0.086	0.493	0.59	0.467

Fig. 12. Fragmento de la matriz de características Ξ_3 para tres dimensiones de algunas de las secuencias de S. *cerevisiae*.

Se calculó la correlación entre todos los elementos de las matrices Ξ y aquellos elementos que mostraron una correlación mayor al 80% fueron eliminados, porque se consideró que contenían información redundante. Así, el número de columnas de la matriz de 2 dimensiones se redujo a 100, el de tres dimensiones a 180 y el de 4 dimensiones a 300. Las secuencias de proteínas quedaron así representadas en el espacio multidimensional Λ.

3.3 Clasificación de las secuencias

Utilizamos mapas autoorganizados (SOMs) para obtener una clasificación no sesgada de todas las secuencias de proteínas de S. *cerevisiae*. Como se mencionó en la introducción, los mapas de Kohonen tienen la característica de ser no supervisados, por lo que no es necesario hacer ninguna suposición *a priori* respecto a la estructura de los *clusters*[5]. La red se va autoorganizando y permite que las características relevantes de los patrones de entrada emerjan durante el proceso de clasificación. Adicionalmente, la disposición bidimensional[6] de las neuronas de salida, permite obtener información útil sobre las relaciones entre los grupos recién descubiertos.

El objetivo de los métodos de clasificación no supervisada es descubrir grupos significativamente distintos entre sí (pero en los cuales los elementos que los integran sean parecidos entre sí) presentes en el conjunto de datos de entrada. Un problema fundamental de estos métodos de clasificación y que ha sido objeto de intensa investigación, es decidir el número óptimo de grupos que se ajuste a nuestro conjunto de datos (Halkidi M et al., 2001). Como el objetivo de estos métodos es descubrir *clusters* que no se conocen con anterioridad, es difícil tener una idea anticipada con respecto al mejor número de clases en las que debemos clasificar nuestro conjunto de datos. Al procedimiento para evaluar los resultados de un algoritmo de *clustering* se le conoce como validación de *clusters* ("cluster validity"). Muchas técnicas han sido propuestas para validar resultados de *clustering* y en general se basan en dos criterios (Berry MJA & Linoff G, 1996):

1) Compactación.- Los miembros de cada *cluster* deben estar lo más cercanos entre sí.

2) Separación.- Los *clusters* mismos deben estar lo más separados posible unos de otros.

Uno de los algoritmos de *clustering* que más se ha trabajado desde la perspectiva de validación de *clusters* es el "Fuzzy c-means" (Bezdek JC, 1974). Se trata de un método de clasificación no supervisada, basado en los principios de la lógica difusa (Zadeh LA, 1965), donde los elementos no pertenecen de manera exclusiva a un *cluster*, sino que su pertenencia a un conjunto específico se da en términos de una función de membresía.

Los métodos de validación propuestos para este algoritmo nos parecieron particularmente apropiados para resolver el problema de establecer el número óptimo de clases en nuestros datos de entrada. Así, para tener una idea aproximada del número de neuronas de salida para los mapas de Kohonen hicimos una exploración previa de nuestros datos utilizando el algoritmo "Fuzzy c-means" (FCM).

3.3.1 Determinación del número óptimo de Clases

El algoritmo FCM puede considerarse como una extensión del algoritmo clásico c-means, para aplicaciones difusas. El objetivo de las técnicas de validación que se han propuesto para este algoritmo es encontrar el esquema de clasificación en el que el mayor número de vectores de entrada exhiban un alto grado de membresía para *clusters* específicos. Para obtener una medida de esto se han

[5] Utilizamos el término *cluster* para hacer referencia a los conjuntos, por extensión de la expresión *clustering* que se ha utilizado para denominar genéricamente a este método de clasificación y para la cual no hay una traducción textual que conserve el significado original. A lo largo de este trabajo utilizaremos *cluster*, conjunto y grupo de manera indistinta.

[6] En la mayoría de las aplicaciones prácticas, aunque, en teoría, los SOMs pueden ser n-dimensionales.

definido varios coeficientes y se han establecido diversos criterios de evaluación de los *clusters* (Rezaee RM. et al., 1998; Windham MP, 1981; Xie XL & Beni G, 1991). Aquí presentaremos los dos coeficientes más comúnmente usados, definidos originalmente por Bezdek (Bezdek JC, 1974) y utilizados en la definición del método conocido coloquialmente como "el criterio del codo": el coeficiente de partición (p_c) y la entropía de la partición (p_e).

Así, el método para establecer el número óptimo de *clusters* C* sería como sigue:

1) Establecer C_{max}, el número máximo de clusters a considerar.

2) Fijar C ← 2.

3) Aplicar el algoritmo FCM, para determinar los centros de los *clusters* difusos y los valores de membresía de los datos de entrada.

4) Calcular el coeficiente de partición (p_c): $p_c = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{C}\mu_{ik}^2$

donde: K= Número de datos u objetos de entrada.
C= Número de *clusters*
μ_{ik}= Valor de membresía del objeto k al *cluster* i.

5) Calcular la entropía de la partición (p_e): $p_e = -\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{C}\mu_{ik}\cdot\ln(\mu_{ik})$

6) Incrementar C ← C+1.

7) Si C ≤ C_{max} regresar al paso 3.

8) Para cada valor de C, graficar los valores de p_c y p_e y encontrar el número óptimo de *clusters* C* mediante la identificación del "codo".

Tanto el coeficiente de partición como la entropía de la partición tienden hacia un comportamiento monótono dependiendo del número de *clusters*. p_c / p_e muestra un patrón de decremento/incremento con respecto a C. La manera de identificar el número óptimo de clusters C* consiste en encontrar el valor de C para el cual la entropía de la partición cae por debajo de la tendencia ascendente y el valor del coeficiente de partición un poco arriba de la tendencia descendente. En la gráfica de la fig.13 se muestra un ejemplo del comportamiento de p_c y p_e y se indica con un círculo el punto donde se cumplen las condiciones antes mencionadas, que corresponde al número óptimo de *clusters*.

De esta forma, para cada una de nuestras matrices de características aplicamos el algoritmo antes descrito fijando C_{max} en 16 y luego graficamos los resultados y obtuvimos el número óptimo de clases para nuestros datos. Este valor lo usamos para establecer los parámetros de configuración de la red de Kohonen, que describiré en la siguiente sección.

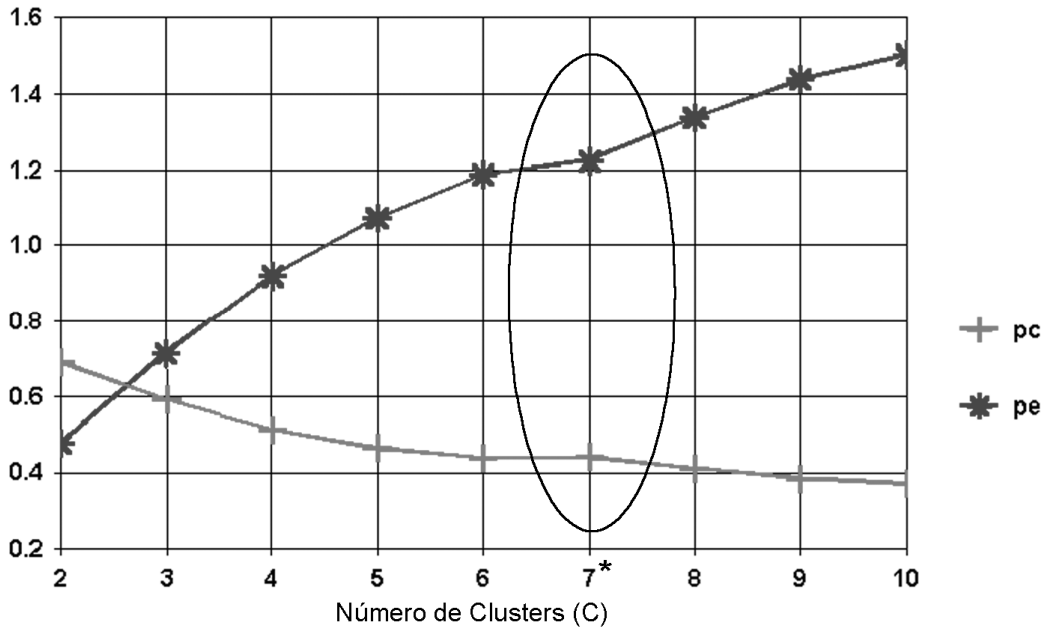


Fig. 13. Método para obtener el número óptimo de *clusters* mediante "el criterio del codo". En este ejemplo, el número óptimo de *clusters* C*=7.
(Imagen tomada del tutoral del programa DataEngine (MIT GmbH, 1997)).

3.3.2 Mapas Autoorganizados

Los SOMs fueron el método que utilizamos para la clasificación de las secuencias de proteínas de S. *cerevisiae*. Ya en la introducción (1.4.2) presentamos este método de clasificación no supervisado y esbozamos el algoritmo de entrenamiento que está implementado en el programa Data Engine (MIT GmbH, 1997), mismo que utilizamos en esta etapa para generar los mapas de Kohonen de nuestras secuencias.

Utilizando como entrada las matrices de vectores de características Ξ_1-Ξ_4, obtenidas según el procedimiento descrito en la sección 3.2 (plano Λ), establecimos los mismos parámetros de inicio para todas las matrices: mapa de dos dimensiones, asignación aleatoria de los valores de inicio de los pesos de las neuronas (entre 0 y 1), semilla unificada y 1000 ciclos de entrenamiento con presentación aleatoria de los vectores de entrada. En todos los casos, se comprobó que al terminar los 1000 ciclos de entrenamiento, la tasa de aprendizaje estuviera muy cerca de 0 (tasa de aprendizaje < $5x10^{-5}$), lo cual nos indica que la topología del mapa de salida ya no está siendo modificado en cada nuevo ciclo de entrenamiento. Con ello verificamos que se ha llegado a la mejor clasificación de los datos de acuerdo con los parámetros de inicio.

El resultado de este proceso de clasificación es una tabla de 0´s y 1´s, de dimensiones NxC, donde N=número de vectores de entrada y C= número de neuronas de salida, en la que cada renglón tiene un sólo 1, que corresponde a la

neurona ganadora para ese vector de entrada específico. En la fig.14 se presenta un fragmento de esta tabla de salida.[7]

	Neur_1-1	Neur_1-2	Neur_1-3	Neur_2-1	Neur_2-2	Neur_2-3	Neur_3-1	Neur_3-2	Neur_3-3
YEL054C	0	0	0	1	0	0	0	0	0
YEL008C-A	0	0	0	0	0	0	1	0	0
YER137C	0	0	0	1	0	0	0	0	0
YOR166C	0	1	0	0	0	0	0	0	0
YOR073W-A	0	0	0	0	0	0	1	0	0
YHL006C	0	0	0	0	0	0	0	0	1
YFL011W	1	0	0	0	0	0	0	0	0
YMR237W	0	0	0	0	0	1	0	0	0
YDL097C	0	0	0	0	1	0	0	0	0
YLR200W	0	0	0	1	0	0	0	0	0
YMR294W-A	0	0	0	0	0	0	0	1	0
YGR066C	0	1	0	0	0	0	0	0	0
YGL049C	0	1	0	0	0	0	0	0	0
YKL160W	0	0	0	0	1	0	0	0	0

Fig. 14. Fragmento de la matriz de salida de una red de Kohonen. Los 1´s en cada renglón indican la pertenencia de ese dato de entrada a una neurona específica. Aquí por ejemplo, el primer elemento (YEL054C), pertenece a la neurona 2-1 del mapa.

De acuerdo con estos resultados, decidimos generar tantas clases de proteínas como neuronas de salida en el SOM, tomando como criterio la neurona ganadora para cada secuencia.

3.3.3 Verificación del método

Dado que el criterio utilizado para la determinación del número óptimo de clases se basa en un algoritmo muy distinto del que se utilizó para hacer la clasificación, hicimos comparaciones entre las clasificaciones obtenidas por ambos métodos, que nos permitieran validar la estrategia utilizada para la determinación del número óptimo de clases.

Con esta finalidad, elaboramos un programa en perl (*find_identity.pl*) que nos permitió identificar genes idénticos entre las clases generadas tanto utilizando el algoritmo FCM, como los mapas de Kohonen y posteriormente analizamos el porcentaje de identidad entre los conjuntos y utilizamos una prueba χ^2 de bondad de ajuste, para determinar si los conjuntos generados por ambos métodos eran significativamente distintos.

Un procedimiento semejante se utilizó para comparar los conjuntos obtenidos con las distintas matrices Ξ_1, Ξ_2, Ξ_3 y Ξ_4 con la finalidad de comprobar si el incremento en la cantidad de información de entrada de la red nos arrojaba clasificaciones distintas.

[7] Cabe aclarar que esta tabla se incluye sólo con la finalidad de explicar la forma en que determinamos la pertenencia de los genes a cada clase y no como parte de los resultados, que se presentarán en la siguiente sección.

3.4 Análisis y validación de la clasificación

Como resultado del proceso de clasificación con SOM´s, obtuvimos 9 conjuntos de genes para cada una de las matrices de entrada (Ξ_1, Ξ_2, Ξ_3 y Ξ_4). El siguiente paso, consistió en analizar los conjuntos de genes obtenidos, para comprobar si la clasificación generada mostraba un significado biológico interesante en términos de la función biológica de las proteínas asociadas a dichos genes. Cabe aclarar que establecemos la correlación directa gen-proteína, debido a que trabajamos con ORF´s, es decir genes que codifican para proteínas. Según el planteamiento que hicimos al comienzo de esta sección, nosotros asumimos que debe existir un mapeo entre la estructura, descrita en el plano matemático Λ y la función de la proteína. Una vez que hemos clasificado nuestro universo de proteínas de acuerdo a la información codificada en este plano matemático, debemos encontrar una base de datos con información funcional, que podamos utilizar para hacer el mapeo de nuestros conjuntos hacia el plano biológico Θ. A este procedimiento, lo llamaremos "mapeo π".

En primer término se hizo un análisis muy general de los conjuntos obtenidos en términos de parámetros básicos y generales (tales como la longitud de la cadena y la composición de aminoácidos) para descartar que nuestra clasificación estuviera reflejando relaciones triviales entre las cadenas de proteínas.

A continuación, y habiendo descartado que se tratara de un clasificador trivial, procedimos a seleccionar la herramienta de validación más adecuada para determinar si las clases obtenidas presentaban alguna correlación con categorías funcionales de las proteínas de *Saccharomyces*.

3.4.1 Mapeo π

De acuerdo con nuestra hipótesis de trabajo, mencionada en el párrafo anterior, esperamos encontrar cierta correlación entre los grupos obtenidos con nuestro clasificador y categorías funcionales de las proteínas. Es en ese sentido que el paso de validación requiere que utilicemos como referencia alguna clasificación funcional de proteínas. Estas clasificaciones generan grupos de proteínas con base en similitud funcional en términos de mecanismos de reacción enzimática, participación en rutas bioquímicas, asociaciones funcionales y localización celular. (Ouzounis CA et al., 2003)

Una de las primeras clasificaciones funcionales es la conocida como EC (Enzyme Commission) (Bairoch, 1994), que es una clasificación jerárquica, que agrupa solamente a las proteínas con función enzimática de acuerdo al tipo de reacción, mecanismos enzimáticos, sustratos sobre los que actúan, etc. Otras bases de datos más generales con información funcional incluyen MIPS (Munich Information Center for Protein Sequence) (Mewes et al., 1999), que esta constituida por una recopilación de anotaciones para varios genomas, WIT/ERGO, (Overbeek et al., 2000) una base de datos comercial con información de funciones conservadas y reconstrucciones metabólicas, STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (Snel et al., 2000) que incluye funciones conservadas, fusiones génicas y perfiles filogenéticos, KEGG (Kyoto Encyclopaedia of Genes and Genomes) (Kanehisa et al., 2002) que es una clasificación muy completa de rutas metabólicas y COG (Clusters of Ortrhologous Groups) (Tatusov et al., 1997) que es una clasificación filogenética de proteínas, por mencionar solo algunas de las más conocidas. Sin embargo, ninguna de ellas se acerca a una clasificación funcional que cubra todas las funciones en todos los organismos. Algunas se centran en

funciones específicas en diferentes especies y otras en todas las funciones para una especie en particular. Por otro lado, en muchas de ellas se mezcla información funcional referente a procesos bioquímicos, con funciones celulares y expresiones fenotípicas (Gerstein, 2000), lo que hace difícil usarlas como herramienta de validación.

Una aproximación más general a la estructura lógica de una clasificación funcional ha sido adoptada por el consorcio Gene Ontology (**GO**) (Ashburner et al., 2000), cuyo objetivo es obtener una clasificación funcional mediante la creación de un diccionario controlado de términos y sus relaciones para describir **función molecular**, **proceso biológico** y **componente celular** de las proteínas. De los esquemas de clasificación funcional propuestos hasta la fecha, GO es el que tiene un mayor potencial para ser utilizado como una herramienta general de referencia para distintas especies y grupos de proteínas (Whisstock JC & Lesk AM, 2003). GO está conformado por tres ontologías independientes:

a) Proceso biológico (**BP**), que hace referencia al objetivo biológico para el cual el gen en cuestión contribuye, está compuesto de una o varias funciones moleculares y frecuentemente involucran una transformación química o física, en el sentido de que algo entra en el proceso y algo diferente resulta de éste (crecimiento y mantenimiento celular, transducción de señales, etc.).

b) Función molecular (**MF**), que puede definirse como la actividad bioquímica de un producto génico (o proteína). Describe solamente lo que ocurre, sin especificar dónde o cuándo tiene lugar (enzima, transportador, ligando, etc.).

c) Componente celular (**CC**), que hace referencia al lugar de la célula en el que el producto génico es activo (ribosoma, núcleo, membrana celular, etc.).

Las ontologías han sido utilizadas desde hace mucho tiempo en un intento por describir todas las entidades dentro de un área de la realidad y todas las relaciones entre esas entidades. Una ontología está compuesta por un conjunto de términos perfectamente definidos y un conjunto de relaciones, también perfectamente definidas. La estructura misma refleja una representación actual del conocimiento biológico y sirve como una guía para la organización de nuevos datos (Gene Ontology Consortium, 2004). Los términos en GO están organizados en estructuras llamadas Gráficas Acíclicas Dirigidas (DAG´s), y difieren de las jerarquías en el sentido de que un “hijo” o término más especializado puede tener muchos “padres” o términos menos especializados (fig.15).

Con base en lo anteriormente descrito y con la ventaja adicional de que GO está particularmente completa para el genoma de *Saccharomyces cerevisiae*, fue que decidimos utilizarla como herramienta de validación de nuestra clasificación.

Aunque existen varias herramientas que permiten encontrar asociaciones entre conjuntos de genes y términos de GO, como el Term Finder (Boyle et al., 2004), preferimos elaborar nuestra propia herramienta, para tener un mayor control sobre los datos y poder generar un índice que nos facilitara las comparaciones. La herramienta que desarrollamos en perl para esta parte del análisis (*valida.pl*), nos permite obtener un conjunto de índices indicativos del grado de sobrelapamiento entre nuestros conjuntos y las categorías funcionales descritas en GO.

El primer paso consistió en obtener los archivos de asociación de S. *cerevisiae* para cada una de las ontologías de GO. Estos archivos contienen el

listado de genes asociados a cada uno de los términos de GO, así como las relaciones entre los distintos términos. Utilizando un algoritmo para la reconstrucción de gráficas, generamos el DAG correspondiente y a continuación analizamos cada uno de los términos, empezando por los hijos, buscando asociaciones entre los genes de cada uno de nuestros conjuntos y los distintos términos de GO en el DAG. Mediante la utilización de un índice, diseñado ex-profeso, asignamos un valor de representatividad a cada término de GO en el DAG y de esta forma pudimos obtener una medida que nos indicara la correlación entre nuestros conjuntos de genes y las categorías funcionales descritas en la ontología.

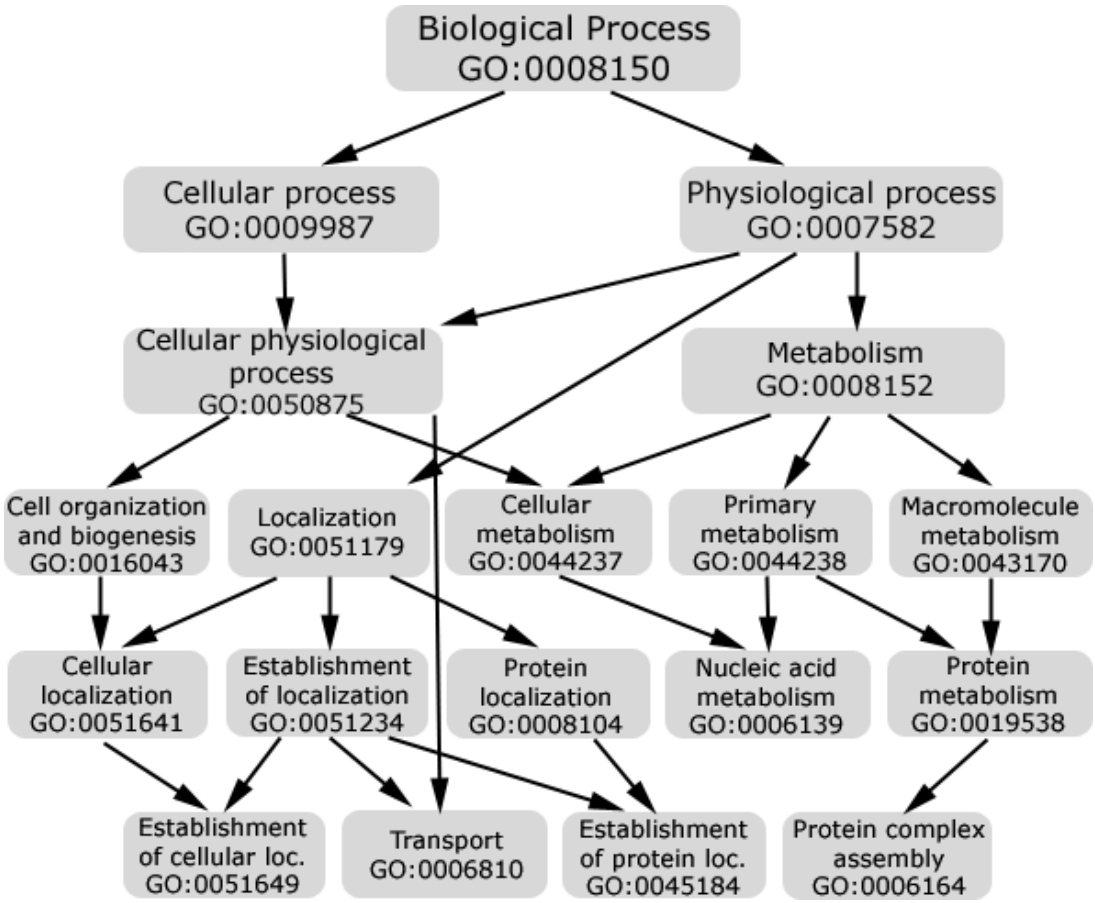


Fig. 15. Fragmento del DAG de GO para la ontología de proceso biológico, donde puede notarse la relación muchos a muchos entre términos “padres” e “hijos”.

Definimos la variable **Q** como un índice de calidad de ajuste de las proteínas de un conjunto en relación a cada término de GO. Si definimos **Ps** como el conjunto de proteínas contenidas en uno de nuestros conjuntos y **Pn** como el conjunto de proteínas asociadas a un término específico de GO en el DAG de S. *cerevisiae*, entonces el subconjunto de proteínas de nuestro conjunto asociadas a un término específico de GO estaría representado por la siguiente expresión: $In=Ps \cap Pn$.

Si denotamos como $|X|$ la cardinalidad del conjunto X, podemos definir dos índices de calidad como sigue:

$\rho = \frac{|In|}{|Pn|}$ Índice de cobertura. Nos indica en qué proporción están representados en nuestro conjunto los genes asociados a un término de GO.

$\omega = \frac{|In|}{|Ps|}$ Índice de representación. Nos indica qué porcentaje de nuestro conjunto se encuentra formando parte de un término de GO.

Para evitar que nuestros índices (ρ y ω) se vean directamente afectados por el tamaño del conjunto, los consideramos como desviaciones de la distribución modelo de la especie. Es decir, calculamos valores esperados para ambos parámetros asumiendo que los genes en nuestro conjunto estuvieran distribuidos de la misma forma en que se distribuyen todos los genes de la especie en cada uno de los términos del DAG y calculamos ρ y ω como las desviaciones respecto a estos valores esperados.

Si definimos P_G como el total de genes anotados para la especie en el DAG de GO, podemos calcular los valores esperados como:

$\rho_E = \frac{|Ps|}{|P_G|}$ Valor esperado para el índice de cobertura, y

$\omega_E = \frac{|Pn|}{|P_G|}$ Valor esperado para el índice de representación.

De aquí, podemos redefinir nuestros estimadores como desviaciones de lo esperado de la siguiente manera:

$\rho_W = \rho - \rho_E$ y $\omega_W = \omega - \omega_E$

Nuestro índice de calidad de ajuste (**Q**) puede incorporar ambos estimadores, si los consideramos como componentes de un vector, de acuerdo con la siguiente expresión:

$$Q = \sqrt{\rho_W^2 + \omega_W^2}$$

Así, tenemos un único valor que nos indica las desviaciones de nuestro conjunto de proteínas respecto a la distribución actual de la especie, para cada término del DAG de GO. Estos valores, cuando son distintos de 0 pueden interpretarse como términos de GO que están sobre-representados o sub-representados en nuestro conjunto.

De esta forma, tenemos ya una medida que nos permite saber si los conjuntos que obtuvimos con nuestro clasificador se asocian de manera específica a ciertas categorías funcionales. Sin embargo, es necesario tener un punto de

referencia que nos permita comparar los valores obtenidos para saber si las desviaciones encontradas tienen un significado estadístico.

Para alcanzar este objetivo, generamos conjuntos aleatorios de tamaños semejantes a nuestros conjuntos y calculamos los valores de Q para cada uno de ellos. Luego, aplicamos una prueba estadística que nos permitiera establecer si los valores de Q encontrados en nuestros conjuntos eran significativamente distintos de los que obtendríamos a partir de conjuntos obtenidos de manera aleatoria.

Por la naturaleza de nuestros datos, fue necesario aplicar una prueba estadística no paramétrica. Decidimos utilizar la “U” de Mann Whitney para comparar las distribuciones de nuestros conjuntos con las que obtuvimos de los conjuntos aleatorios. Estas comparaciones nos permitieron llegar a conclusiones importantes en relación a nuestros conjuntos, así como intentar discriminar entre las clasificaciones obtenidas a partir de las matrices Ξ_1, Ξ_2, Ξ_3 y Ξ_4, para determinar la mejor representación de las secuencias.

4

RESULTADOS

En esta sección presentaré las tablas y gráficas que resumen los resultados obtenidos para cada etapa, resaltando los aspectos más relevantes de cada una de ellas. El mapeo de las secuencias (4.1), la determinación del número óptimo de clases (4.2.1), la clasificación mediante mapas Autoorganizados (4.2.2), la verificación del método para la determinación del número de clases (4.2.3), el análisis y validación de las clases obtenidas (4.3) y por último, se hace una exploración de las potencialidades del método como clasificador de nuevas secuencias (4.4).

4.1 Mapeo λ

De acuerdo con el procedimiento descrito en la sección (3.2.2), obtuvimos las matrices de características para las 6700 secuencias de aminoácidos del conjunto completo de ORF´s de S. *cerevisiae* en una, dos, tres y cuatro dimensiones. Las fig.16-19 muestran fragmentos de estas matrices, después de haberse eliminado las columnas que mostraron altas correlaciones.

Genes	f_A	X_{1A}	σ_{X1A}	f_C	X_{1C}	σ_{X1C}	f_D	X_{1D}	σ_{X1D}	f_E	X_{1E}	σ_{X1E}	f_F	X_{1F}	σ_{X1F}
YEL054C	0.889	0.425	0.207	0.056	0.855	0	0.444	0.573	0.214	0.667	0.606	0.29	0.222	0.491	0.315
YER137C	0.364	0.62	0.244	0.182	0.664	0.325	0.318	0.393	0.215	0.545	0.414	0.207	0.091	0.598	0.26
YOR166C	0.419	0.477	0.226	0.163	0.478	0.177	0.814	0.423	0.2	0.744	0.533	0.295	0.512	0.661	0.211
YFL011W	0.604	0.528	0.218	0.226	0.518	0.239	0.264	0.374	0.255	0.396	0.552	0.248	0.774	0.583	0.234
YMR237W	0.519	0.506	0.238	0.123	0.626	0.216	0.605	0.529	0.275	0.543	0.532	0.198	0.383	0.522	0.277
YDL097C	0.526	0.467	0.226	0.105	0.583	0.157	0.491	0.517	0.252	0.614	0.468	0.28	0.281	0.516	0.219
YGR066C	0.355	0.549	0.303	0.194	0.669	0.237	0.581	0.507	0.24	0.806	0.556	0.262	0.516	0.569	0.227
YGL049C	0.539	0.508	0.255	0.056	0.733	0.044	0.584	0.568	0.211	0.944	0.505	0.222	0.303	0.601	0.239
YDR434W	0.379	0.53	0.264	0.061	0.416	0.242	0.5	0.505	0.212	0.515	0.594	0.277	0.424	0.494	0.261
YOL083W	0.275	0.49	0.296	0.175	0.391	0.203	0.475	0.458	0.3	0.9	0.54	0.202	0.725	0.458	0.291
YPL041C	0.143	0.671	0.157	0.086	0.493	0.293	0.2	0.488	0.148	0.171	0.46	0.189	0.457	0.486	0.231
YHR023W	0.319	0.528	0.261	0.061	0.481	0.25	0.467	0.536	0.247	0.891	0.561	0.239	0.266	0.375	0.197
YLR201C	0.433	0.575	0.217	0.067	0.027	0.015	0.3	0.643	0.121	0.6	0.505	0.229	0.467	0.493	0.265
YOL149W	0.345	0.246	0.257	0.069	0.669	0.115	0.483	0.525	0.223	0.483	0.614	0.3	0.276	0.418	0.284

Fig. 16. Fragmento de la matriz Ξ_1.

Genes	f_A	X_{1A}	σ_{X1A}	X_{2A}	σ_{X2A}	f_C	X_{1C}	σ_{X1C}	X_{2C}	σ_{X2C}	f_D	X_{1D}	σ_{X1D}	X_{2D}	σ_{X2D}
YEL054C	0.889	0.425	0.207	0.519	0.238	0.056	0.855	0	0.846	0	0.444	0.573	0.214	0.394	0.221
YER137C	0.364	0.62	0.244	0.558	0.231	0.182	0.664	0.325	0.558	0.202	0.318	0.393	0.215	0.615	0.11
YOR166C	0.419	0.477	0.226	0.593	0.265	0.163	0.478	0.177	0.675	0.23	0.814	0.423	0.2	0.497	0.242
YFL011W	0.604	0.528	0.218	0.544	0.27	0.226	0.518	0.239	0.542	0.167	0.264	0.374	0.255	0.518	0.259
YMR237W	0.519	0.506	0.238	0.486	0.286	0.123	0.626	0.216	0.496	0.17	0.605	0.529	0.275	0.504	0.252
YDL097C	0.526	0.467	0.226	0.457	0.227	0.105	0.583	0.157	0.714	0.206	0.491	0.517	0.252	0.408	0.205
YGR066C	0.355	0.549	0.303	0.545	0.185	0.194	0.669	0.237	0.685	0.13	0.581	0.507	0.24	0.509	0.251
YGL049C	0.539	0.508	0.255	0.517	0.218	0.056	0.733	0.044	0.426	0.201	0.584	0.568	0.211	0.506	0.244
YDR434W	0.379	0.53	0.264	0.55	0.236	0.061	0.416	0.242	0.51	0.198	0.5	0.505	0.212	0.504	0.292
YOL083W	0.275	0.49	0.296	0.424	0.264	0.175	0.391	0.203	0.667	0.272	0.475	0.458	0.3	0.519	0.241
YPL041C	0.143	0.671	0.157	0.467	0.24	0.086	0.493	0.293	0.467	0.178	0.2	0.488	0.148	0.448	0.207
YHR023W	0.319	0.528	0.261	0.511	0.215	0.061	0.481	0.25	0.638	0.264	0.467	0.536	0.247	0.552	0.25
YLR201C	0.433	0.575	0.217	0.57	0.279	0.067	0.027	0.015	0.412	0.235	0.3	0.643	0.121	0.386	0.292
YOL149W	0.345	0.246	0.257	0.456	0.173	0.069	0.669	0.115	0.656	0.344	0.483	0.525	0.223	0.504	0.193

Fig. 17. Fragmento de la matriz Ξ_2.

Genes	f_A	X_{1A}	σ_{X1A}	X_{2A}	σ_{X2A}	X_{3A}	σ_{X3A}	Y_{3A}	σ_{Y3A}	f_C	X_{1C}	σ_{X1C}	X_{2C}	σ_{X2C}	X_{3C}
YEL054C	0.889	0.425	0.207	0.519	0.238	0.688	0.227	0.625	0.25	0.056	0.855	0	0.846	0	0.5
YER137C	0.364	0.62	0.244	0.558	0.231	0.792	0.177	0.6	0.2	0.182	0.664	0.325	0.558	0.202	0.625
YOR166C	0.419	0.477	0.226	0.593	0.265	0.688	0.236	0.507	0.258	0.163	0.478	0.177	0.675	0.23	0.679
YFL011W	0.604	0.528	0.218	0.544	0.27	0.503	0.229	0.504	0.27	0.226	0.518	0.239	0.542	0.167	0.611
YMR237W	0.519	0.506	0.238	0.486	0.286	0.553	0.262	0.593	0.228	0.123	0.626	0.216	0.496	0.17	0.589
YDL097C	0.526	0.467	0.226	0.457	0.227	0.55	0.245	0.525	0.245	0.105	0.583	0.157	0.714	0.206	0.458
YGR066C	0.355	0.549	0.303	0.545	0.185	0.545	0.288	0.61	0.276	0.194	0.669	0.237	0.685	0.13	0.738
YGL049C	0.539	0.508	0.255	0.517	0.218	0.569	0.229	0.56	0.305	0.056	0.733	0.044	0.426	0.201	0.64
YDR434W	0.379	0.53	0.264	0.55	0.236	0.64	0.281	0.575	0.302	0.061	0.416	0.242	0.51	0.198	0.444
YOL083W	0.275	0.49	0.296	0.424	0.264	0.489	0.198	0.489	0.174	0.175	0.391	0.203	0.667	0.272	0.554
YPL041C	0.143	0.671	0.157	0.467	0.24	0.767	0.24	0.7	0.24	0.086	0.493	0.293	0.467	0.178	0.667
YHR023W	0.319	0.528	0.261	0.511	0.215	0.548	0.254	0.57	0.25	0.061	0.481	0.25	0.638	0.264	0.516
YLR201C	0.433	0.575	0.217	0.57	0.279	0.604	0.189	0.648	0.313	0.067	0.027	0.015	0.412	0.235	0.5
YOL149W	0.345	0.246	0.257	0.456	0.173	0.629	0.257	0.317	0.15	0.069	0.669	0.115	0.656	0.344	0.571

Fig. 18. Fragmento de la matriz Ξ_3.

Genes	f_A	X_{1A}	σ_{X1A}	X_{2A}	σ_{X2A}	X_{3A}	σ_{X3A}	Y_{3A}	σ_{Y3A}	X_{4A}	σ_{X4A}	Y_{4A}	σ_{Y4A}	Z_{4A}	σ_{Z4A}
YEL054C	0.889	0.425	0.207	0.519	0.238	0.688	0.227	0.625	0.25	0.531	0.258	0.625	0.266	0.531	0.203
YER137C	0.364	0.62	0.244	0.558	0.231	0.792	0.177	0.6	0.2	0.813	0.188	0.531	0.227	0.438	0.141
YOR166C	0.419	0.477	0.226	0.593	0.265	0.688	0.236	0.507	0.258	0.622	0.286	0.533	0.222	0.578	0.225
YFL011W	0.604	0.528	0.218	0.544	0.27	0.503	0.229	0.504	0.27	0.6	0.25	0.519	0.219	0.544	0.244
YMR237W	0.519	0.506	0.238	0.486	0.286	0.553	0.262	0.593	0.228	0.556	0.251	0.648	0.231	0.576	0.247
YDL097C	0.526	0.467	0.226	0.457	0.227	0.55	0.245	0.525	0.245	0.593	0.274	0.593	0.248	0.553	0.186
YGR066C	0.355	0.549	0.303	0.545	0.185	0.545	0.288	0.61	0.276	0.618	0.274	0.568	0.269	0.682	0.223
YGL049C	0.539	0.508	0.255	0.517	0.218	0.569	0.229	0.56	0.305	0.622	0.24	0.583	0.222	0.479	0.244
YDR434W	0.379	0.53	0.264	0.55	0.236	0.64	0.281	0.575	0.302	0.552	0.226	0.608	0.265	0.44	0.198
YOL083W	0.275	0.49	0.296	0.424	0.264	0.489	0.198	0.489	0.174	0.709	0.192	0.855	0.185	0.655	0.258
YPL041C	0.143	0.671	0.157	0.467	0.24	0.767	0.24	0.7	0.24	0.75	0.1	0.75	0.1	0.65	0.22
YHR023W	0.319	0.528	0.261	0.511	0.215	0.548	0.254	0.57	0.25	0.571	0.258	0.593	0.26	0.571	0.25
YLR201C	0.433	0.575	0.217	0.57	0.279	0.604	0.189	0.648	0.313	0.538	0.18	0.519	0.18	0.558	0.263
YOL149W	0.345	0.246	0.257	0.456	0.173	0.629	0.257	0.317	0.15	0.725	0.235	0.525	0.135	0.425	0.21

Fig. 19. Fragmento de la matriz Ξ_4.

4.2 Clasificación de la secuencias

4.2.1 Determinación del número óptimo de clases (*clusters*)

Como mencionamos en la sección (3.3.1), utilizamos la técnica de “el codo” del algoritmo fuzzy-c jeans (FCM) para determinar el número óptimo de *clusters* (En esta sección se utilizan indistintamente los términos clase y *cluster*). En la fig.20 se muestran las gráficas del coeficiente de partición (p_c) y la entropía de la partición (p_e) para algunas de nuestras matrices Ξ.

En las gráficas se muestra en un círculo el primer punto en donde hay un cambio en la tendencia de los índices, que corresponde al número óptimo de *clusters* y se asocia con el valor de 9. Si bien, en todos los casos se presentaron otros “codos” asociados a un mayor número de *clusters*, decidimos utilizar como criterio la primera aparición de “el codo”, que además coincidió en el valor de 9 para todas las matrices. La determinación de este parámetro es importante, ya que debe proporcionarse como parte de la configuración de la red de Kohonen.

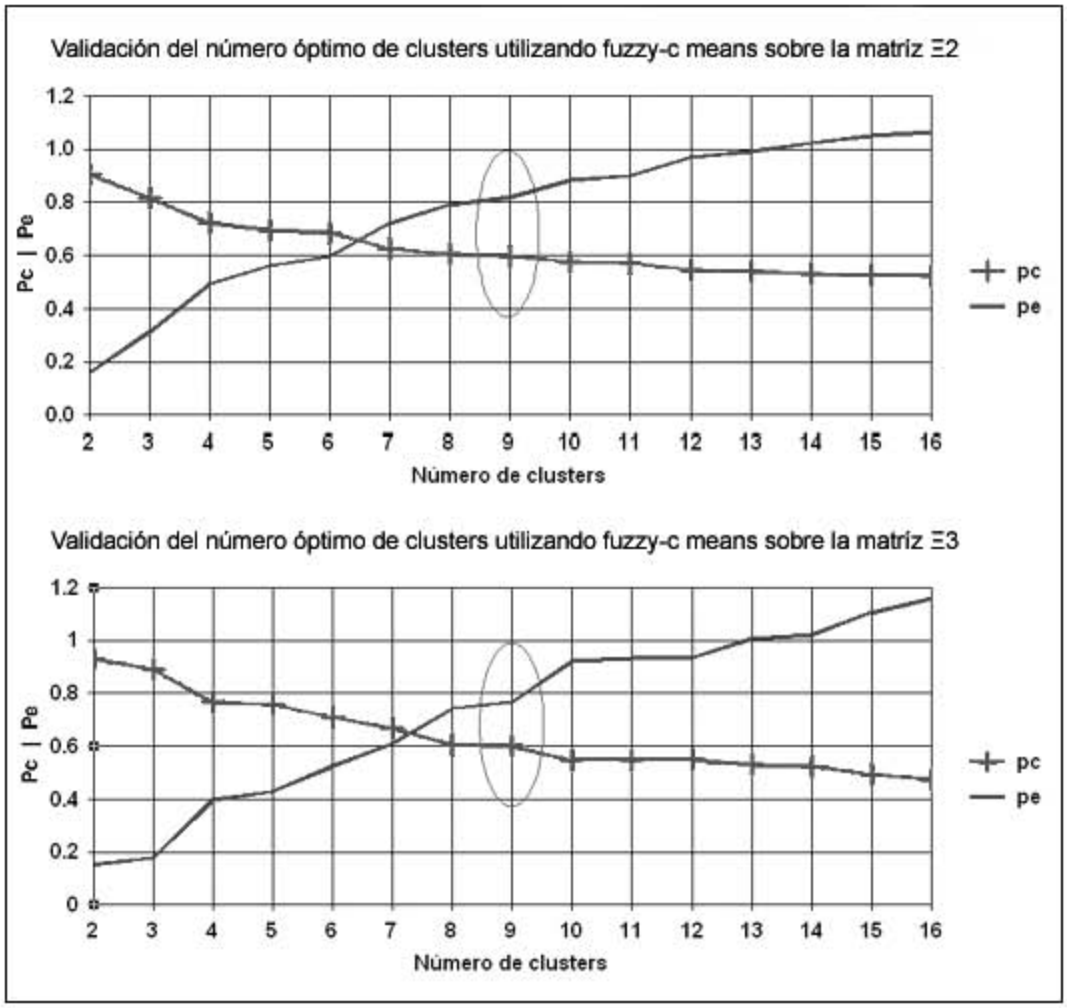


Fig. 20. Gráficas del coeficiente de partición (p_c) y la entropía de la partición (p_e) con respecto al número de *clusters*, utilizando el algoritmo FCM para clasificar las secuencias representadas mediante las matrices Ξ_2 y Ξ_3, respectivamente.

4.2.2 Clasificación con Mapas Autoorganizados (SOMs).

El siguiente paso consistió en entrenar una red de Kohonen con cada una de las matrices Ξ_1-Ξ_4. Los parámetros de la red fueron los mismos para todas las matrices: se definieron mapas de dos dimensiones, con 3 neuronas por dimensión, los vectores de pesos se inicializaron de forma aleatoria con la misma semilla y se ejecutaron 1000 ciclos de entrenamiento, presentando las muestras de manera aleatoria en cada nuevo ciclo. En todos los casos, la tasa de aprendizaje fue menor a 4.5×10^{-5} al terminar los 1000 ciclos de entrenamiento, lo cual nos indica que la topología del mapa de salida ya no está siendo modificado en cada nuevo ciclo de entrenamiento.

Como resultado del entrenamiento de la red, obtuvimos matrices de 0´s y 1´s, análogas a la que se muestra en la fig.14, de 6700 renglones (correspondientes a cada una de las proteínas) por 9 columnas (correspondientes a cada una de las neuronas de salida), en la que cada renglón presenta valores de cero para todas las neuronas, excepto para la neurona ganadora, que tiene asignado un valor de 1.

Agrupando todos los genes que quedaron asociados a cada una de las nueve neuronas de salida, pudimos obtener 9 conjuntos (*sets*[8]) de proteínas para cada una de las 4 matrices **Ξ** analizadas.

4.2.3 Verificación del método

De acuerdo con lo que mencioné en la sección (3.3.3), se hicieron comparaciones entre los conjuntos generados utilizando el algoritmo FCM y los que se obtuvieron con los SOMs, para respaldar la utilización de un algoritmo diferente en la determinación del número óptimo de *clusters*. El razonamiento sobre el que se basa este argumento es que si los conjuntos generados por ambos métodos son muy semejantes, entonces es válida la extrapolación del método que nos permite definir el número óptimo de *clusters* entre un algoritmo y otro.

Hasta ahora hemos hablado del algoritmo FCM sólo en términos del método de validación para determinar el número óptimo de *clusters*, pero no se ha explicado la forma en la que se generan los conjuntos mismos. Como ya mencioné en la sección (3.3) el algoritmo FCM es un método de clasificación difuso, en el sentido de que el proceso de clasificación de las muestras no genera clases perfectamente delimitadas (en las que cada elemento de la muestra puede pertenecer solamente a una de las clases). En los métodos de clasificación difusos, a cada muestra se le asigna un índice de membresía (v), que indica el grado de pertenencia de la muestra al conjunto en cuestión.

Cuando clasificamos un grupo de muestras mediante este algoritmo, el resultado que obtenemos es una matriz de índices de membresía de dimensionalidad NxM, donde N=número de muestras de entrada (6700 en nuestro caso) y M=número de conjuntos generados, donde $0 \le \nu \le 1$. Para cada muestra N de la matriz $\sum_{i=1}^{M} \nu_i = 1$. Valores de ν cercanos a 1 indican un alto grado de pertenencia a ese conjunto.

	1	2	3	4	5	6	7	8	9
YEL054C	0.023041	0.005142	0.001014	0.0123	**0.936549**	0.00783	0.001959	0.011887	0.000278
YER137C	0.000842	0.004012	0.001141	0.002055	**0.980025**	0.002902	0.002003	0.006615	0.000406
YOR166C	0.059271	0.002948	3.20E-06	**0.875737**	2.90E-05	0.005248	1.68E-05	0.056748	1.82E-08
YOR073W-A	0.006083	0.039186	0.011104	0.008834	0.011967	0.030452	0.256806	0.018477	**0.61709**
YFL011W	**0.694354**	0.002723	7.61E-06	0.025127	6.12E-05	0.269224	0.000653	0.007849	3.92E-07
YDL097C	0.013255	0.067534	0.000172	0.064589	0.000192	0.119558	0.00449	**0.730209**	7.59E-07
YMR294W-A	0.004504	0.012082	**0.943906**	0.005485	0.006691	0.007228	0.011719	0.006617	0.001768
YGR066C	0.119269	0.031433	1.47E-05	0.345749	3.54E-05	0.147255	0.000864	0.35538	9.74E-08
YGL049C	0.082221	0.008604	2.48E-05	**0.854271**	5.04E-05	0.00102	5.53E-06	0.053804	1.85E-08
YHR023W	0.001358	0.006471	3.37E-06	0.242171	3.09E-05	0.002492	1.02E-05	**0.747464**	9.47E-09
YDR100W	0.011957	0.017692	0.007291	0.011246	**0.785419**	0.060788	0.071465	0.020078	0.014063
YNL112W	0.180217	0.200485	0.00019	0.102705	0.000127	0.294157	0.002284	0.219835	6.18E-07
YLR157C-A	0.000969	**0.977535**	5.23E-05	0.007192	1.36E-05	0.005988	0.000523	0.007727	2.76E-07
YLR157C-B	0.006241	0.01244	1.97E-07	**0.965547**	1.90E-07	0.001176	5.07E-07	0.014595	1.60E-10
YLR157C-C	0.024279	0.080073	0.072124	0.024187	0.037619	0.073025	0.388642	0.046912	0.253138
YEL055C	0.000287	0.011168	3.69E-07	0.010644	4.16E-06	0.029622	6.87E-05	**0.948205**	3.73E-09

Fig. 21. Matriz de índices de membresía obtenida mediante el algoritmo FCM para la clasificación de las secuencias de proteínas en 9 conjuntos difusos.

[8] A partir de este punto usaré indistintamente los términos clase y set, esto debido a que el etiquetamiento de los conjuntos tanto en tablas como en gráficas se simplifica utilizando el término inglés set.

La fig.21 muestra un fragmento de una de las matrices de salida que obtuvimos como resultado de la clasificación de nuestro conjunto de proteínas con el algoritmo FCM.

Con la finalidad de poder hacer comparaciones entre los conjuntos generados por este método (FCM) y los obtenidos mediante redes de Kohonen (K), establecimos un límite para el valor del índice de membresía que nos permitiera asignar las muestras de manera inequívoca a un solo conjunto. Así, solamente las muestras en las que $\nu \geq 0.5$ fueron asignadas a uno de los conjuntos, quedando sin clasificar todas aquellas en las que no había una dominancia clara en la pertenencia a alguno de los conjuntos. Tal vez sería válido considerar como miembros de un conjunto a todos aquellos genes que mostraran un índice de membresía claramente mayor para el conjunto en cuestión, aunque no necesariamente mayor de 0.5. Sin embargo, para evitar tener que estar discriminando entre valores, a veces muy cercanos entre sí y como nuestro objetivo era establecer si el grueso de los genes de un conjunto eran comunes entre ambos métodos, decidimos mantener este criterio. En la fig.21 se muestran resaltados en negritas los valores que cumplen con la condición antes descrita y que sirvieron para asignar las muestras a alguno de los conjuntos.

Aquí me parece importante aclarar los motivos por los cuales no utilizamos FCM como método de clasificación de nuestras secuencias, que a primera vista podría parecer lo más natural, dado que es un método que nos permite definir el número de clases con un criterio bastante objetivo.

1) La principal razón es que las redes de Kohonen ofrecen posibilidades asociadas con el mapeo bidimensional de los conjuntos que no tenemos en FCM. En particular, una vez que hemos entrenado y etiquetado la red para un mapa de ciertas dimensiones, nos permite hacer una subclasificación de nuestros datos, incrementando las dimensiones del mapa y por consiguiente el número de neuronas de salida y observar de manera gráfica la pertenencia de las nuevas neuronas a las clases previamente etiquetadas, como se muestra en la fig.22.

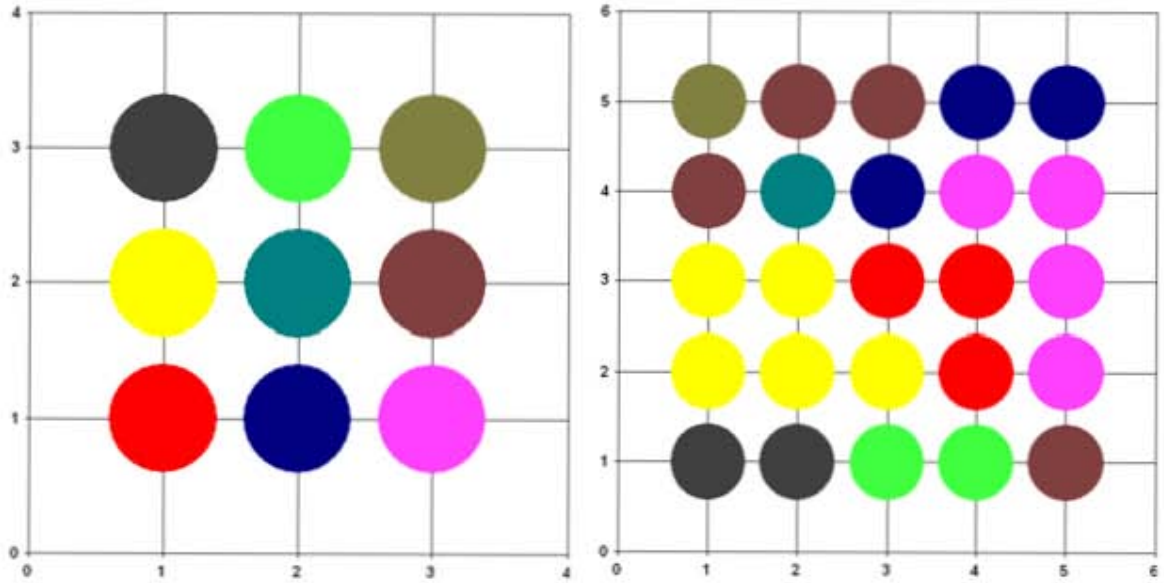


Fig. 22. A la izquierda se muestra un mapa de Kohonen de 3x3 neuronas, etiquetado asumiendo una neurona por clase. A la derecha se muestra un mapa de 5x5 neuronas, obtenido haciendo una subclasificación de los mismos datos y etiquetado bajo la misma definición del primero (obtenido mediante el programa Data Engine(MIT GmbH, 1997)).

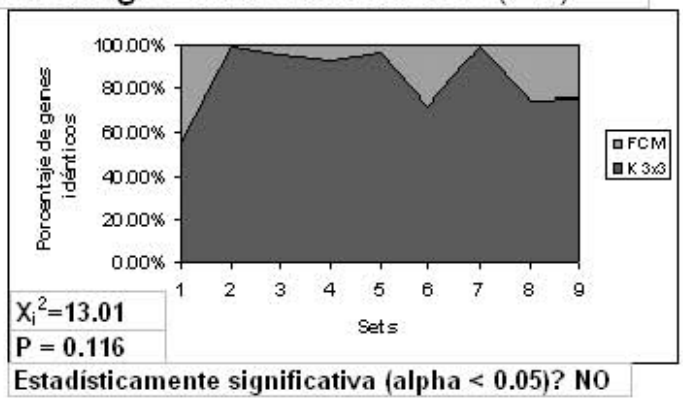
De esta forma podemos ir generando una clasificación jerárquica de todas nuestras secuencias en conjuntos cada vez más específicos.

2) Por otro lado, el algoritmo FCM tiene en este caso la particularidad de que no nos permite establecer clases perfectamente definidas, lo que hubiera ocasionado que muchas de nuestras secuencias quedaran sin poder asignarse a una clase y por lo mismo, a una categoría funcional específica.

El siguiente paso para hacer las comparaciones consistió en la elaboración de un programa (*find_identity.pl*) que nos permitió comparar cada uno de los conjuntos obtenidos mediante FCM con todos los conjuntos generados con SOMs para los mismos datos de entrada y que, como resultado, nos arroja el número de genes idénticos entre cada par de conjuntos comparados. En la fig.23 se muestran los resultados de las comparaciones entre los conjuntos generados con FCM para 9 *clusters* y los conjuntos obtenidos mediante mapas autoorganizados con nueve neuronas de salida para la cuatro matrices Ξ_1-Ξ_4.

Fuzzy-c means clustering 9 Vs Kohonen 3x3 (Ξ1)

	FCM	K 3x3	
	#genes	#genes idéntico	
set1	529	295	56.00%
set2	612	606	99.00%
set3	910	878	96.00%
set4	294	273	93.00%
set5	1122	1092	97.00%
set6	435	314	72.00%
set7	304	300	99.00%
set8	566	426	75.00%
set9	896	679	76.00%
			84.78%



Fuzzy-c means clustering 9 Vs Kohonen 3x3 (Ξ2)

	FCM	K 3x3	
	#genes	#genes idéntico	
set1	247	238	96.00%
set2	327	325	99.00%
set3	561	552	98.00%
set4	1099	1049	95.00%
set5	447	445	100.00%
set6	847	767	91.00%
set7	445	439	99.00%
set8	957	946	99.00%
set9	521	517	99.00%
			97.33%

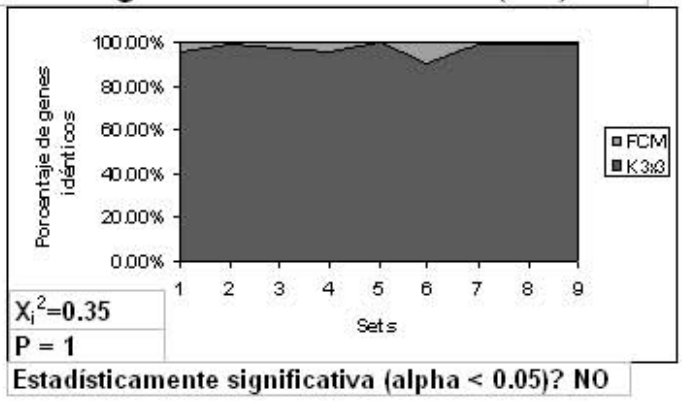
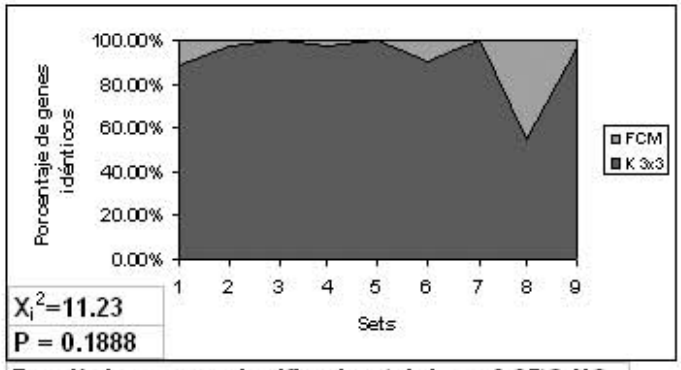


Fig. 23a. Tablas de datos, gráficas y valor de χ^2, de las comparaciones entre los conjuntos de proteínas generados mediante el algoritmo FCM y los conjuntos obtenidos utilizando mapas de Kohonen para las matrices Ξ_1 y Ξ_2.

Fuzzy-c means clustering 9 Vs Kohonen 3x3 (Ξ3)

	FCM	K 3x3	
	#genes	#genes idéntico	
set1	951	851	89.00%
set2	485	477	98.00%
set3	497	497	100.00%
set4	929	912	98.00%
set5	516	516	100.00%
set6	411	372	91.00%
set7	234	234	100.00%
set8	619	339	55.00%
set9	177	171	97.00%
			92.00%



Fuzzy-c means clustering 9 Vs Kohonen 3x3 (Ξ4)

	FCM	K 3x3	
	#genes	#genes idéntico	
set1	750	540	72.00%
set2	404	384	95.00%
set3	959	956	100.00%
set4	198	198	100.00%
set5	120	119	99.00%
set6	384	384	100.00%
set7	482	475	99.00%
set8	543	543	100.00%
set9	229	225	98.00%
			95.89%

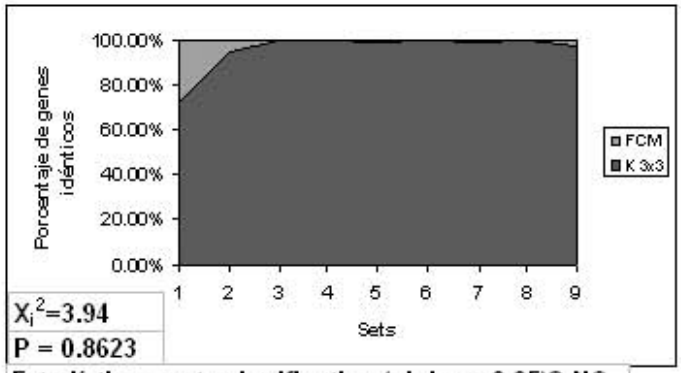


Fig. 23b. Tablas de datos, gráficas y valor de χ^2, de las comparaciones entre los conjuntos de proteínas generados mediante el algoritmo FCM y los conjuntos obtenidos utilizando mapas de Kohonen para las matrices Ξ_3 y Ξ_4.

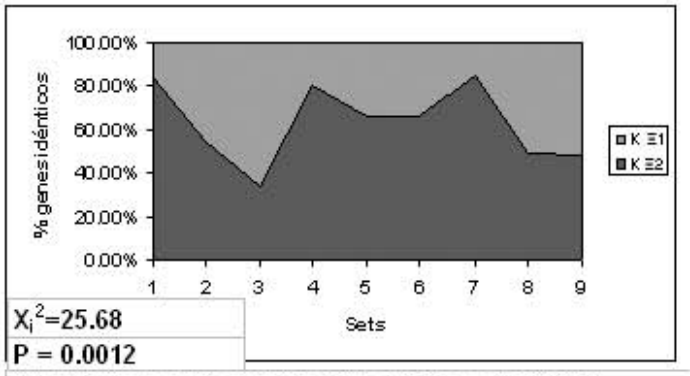
Como puede verse de la fig.23, en todos los casos, los conjuntos generados mediante el algoritmo FCM mostraron una gran similitud con conjuntos específicos generados mediante mapas autoorganizados. El porcentaje de genes idénticos entre ambos métodos resultó ser de 85% para el caso de la matriz Ξ_1 y superior al 90% para las otras tres matrices de entrada. En todos los casos, el valor de χ^2 obtenido nos indica que las distribuciones de ambos grupos de conjuntos no son diferentes entre sí.

Estos datos nos permiten sostener que sí es apropiado haber utilizado la técnica del "codo" para la determinación del número óptimo de *clusters*, ya que con cualquiera de los dos métodos nuestras secuencias de proteínas se agrupan de manera muy semejante.

Ahora bien, otra pregunta que es importante plantearse sabiendo lo anterior es si los conjuntos obtenidos a partir de las matrices Ξ_1, Ξ_2, Ξ_3 y Ξ_4 son diferentes entre sí, o lo que es lo mismo, si al aumentar la dimensionalidad de los vectores de características, estamos agregando información a nuestro clasificador. Para contestarla, hicimos comparaciones entre los conjuntos obtenidos con mapas autoorganizados a partir de las matrices Ξ_1-Ξ_4. En la fig.24 se muestran los resultados de estas comparaciones.

Kohonen 3x3 Ξ1 Vs Ξ2

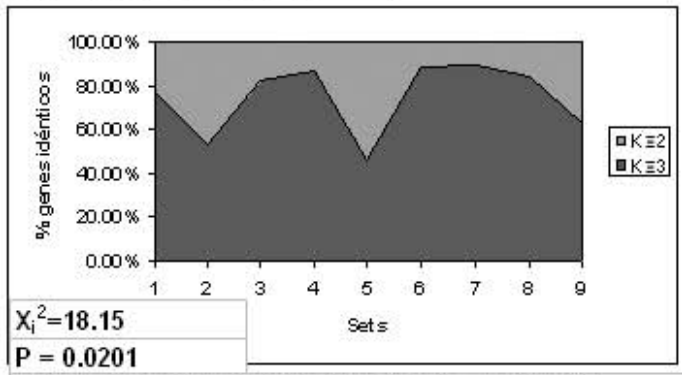
	K 3x3 Ξ1	K 3x3 Ξ2	
	#genes	#genes idénticos	
set1	673	563	84.00%
set2	573	309	54.00%
set3	357	120	34.00%
set4	1387	1103	80.00%
set5	922	608	66.00%
set6	699	459	66.00%
set7	1102	935	85.00%
set8	604	299	50.00%
set9	374	178	48.00%
			63.00%



X_i^2=25.68
P = 0.0012
Estadísticamente significativa (alpha < 0.05)? SI

Kohonen 3x3 Ξ2 Vs Ξ3

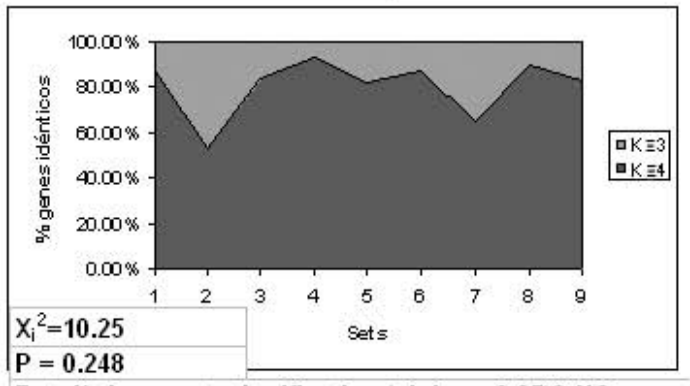
	K 3x3 Ξ2	K 3x3 Ξ3	
	#genes	#genes idénticos	
set1	1081	829	77.00%
set2	721	380	53.00%
set3	537	445	83.00%
set4	1319	1144	87.00%
set5	1026	477	46.00%
set6	598	535	89.00%
set7	524	474	90.00%
set8	557	469	84.00%
set9	328	206	63.00%
			74.67%



X_i^2=18.15
P = 0.0201
Estadísticamente significativa (alpha < 0.05)? SI

Kohonen 3x3 Ξ3 Vs Ξ4

	K 3x3 Ξ3	K 3x3 Ξ4	
	#genes	#genes idénticos	
set1	1115	973	87.00%
set2	1445	768	53.00%
set3	612	517	84.00%
set4	592	551	93.00%
set5	610	502	82.00%
set6	1017	886	87.00%
set7	221	143	65.00%
set8	544	490	90.00%
set9	535	446	83.00%
			80.44%



X_i^2=10.25
P = 0.248
Estadísticamente significativa (alpha < 0.05)? NO

Fig. 24. Tablas de datos, gráficas y valor de χ^2, de las comparaciones entre los conjuntos de proteínas generados utilizando redes de Kohonen, para las matrices Ξ_1, Ξ_2, Ξ_3 y Ξ_4.

Como puede observarse de la fig.24, el porcentaje de identidad entre los conjuntos obtenidos utilizando mapas autoorganizados varía del 63%, para el caso de la comparación de Ξ_1 y Ξ_2, hasta 80% para la comparación entre Ξ_3 y Ξ_4. El resultado de χ^2 mostró que las distribuciones de genes idénticos entre Ξ_1 y Ξ_2 y Ξ_2 y

Ξ_3 sí difieren entre sí, mientras que las distribuciones Ξ_3 y Ξ_4 no difieren significativamente. De aquí, podemos concluir que las clasificaciones obtenidas a partir de las cuatro matrices de entrada no son iguales entre sí y por lo tanto, tiene sentido continuar el análisis de las cuatro clasificaciones hasta la etapa de validación, en la que podríamos proponer una de ellas como la que mejor se ajusta a categorías funcionales y por lo tanto, a nuestra hipótesis de trabajo.

4.3 Análisis y validación de la clasificación

4.3.1 Estadística de los conjuntos

Como mencionamos en la sección (3.4), el primer paso de análisis de las clasificaciones obtenidas, consistió en hacer una exploración de los conjuntos en términos de parámetros estadísticos, tales como longitud de las secuencias y composición de aminoácidos, para descartar que nuestro clasificador se esté basando en descriptores triviales de las cadenas de proteínas. Con este fin, hicimos programas que obtienen estos descriptores a partir de archivos con las secuencias completas de todas las proteínas agrupadas en cada una de las clases obtenidas (*longitud.pl* y *frecuencia.pl*).

En esta sección se mostrarán los resultados obtenidos para el análisis de longitud de la cadena y frecuencia de aminoácidos para los conjuntos obtenidos a partir de cada una de las matrices de entrada (Ξ_1-Ξ_4**)**. Las figs.25 y 26 corresponden al análisis de Ξ_1**,** las figs.27 y 28 al de Ξ_2, las figs.29 y 30 a Ξ_3 y por último, las figs. 31 y 32 corresponden al análisis de Ξ_4.

Se obtuvo la longitud promedio de las cadenas de aminoácidos de todas las proteínas asociadas a cada conjunto y se calculó la desviación estándar de esta medida. Se calculó el *Z-score* (que nos indica la distancia en términos de número de desviaciones estándar entre dos valores) entre todos los conjuntos para determinar si había una separación clara entre éstos debida únicamente a la longitud de la cadena. Los resultados para Ξ_1 (fig.25) muestran, que aunque algunos de los conjuntos, en particular en este caso, el 2 y el 9 sí tienen cadenas de longitud distinta a casi todos los demás, no son diferentes entre sí, y en el caso del 9, tampoco es diferente de los conjuntos 2 y 3. Se calculó también este parámetro para la totalidad de las proteínas de la especie y también en este caso, sólo los conjuntos 2 y 9 mostraron diferencias con respecto a la longitud global de la especie. En la tabla de *Z-score* de la fig.25 se muestran resaltados en negritas los valores mayores a 2 desviaciones estándar.

Para el siguiente análisis se calculó la frecuencia de todos los aminoácidos en cada una de las cadenas de los conjuntos y a partir de este valor, se obtuvo una frecuencia promedio y una desviación estándar de la frecuencia de cada aminoácido en cada conjunto. En la tabla A2.I del apéndice 2 se muestran los valores de frecuencia promedio y desviación estándar de todos los aminoácidos en cada conjunto. La fig.26 muestra las gráficas generadas a partir de esta información para el caso de Ξ_1. En particular, en la segunda gráfica, se ordenaron los aminoácidos según su frecuencia en la especie. Puede notarse que aunque efectivamente hay fluctuaciones en la frecuencia de los distintos aminoácidos entre conjuntos, algunos de los cuales muestran picos de mayor o menor frecuencia respecto a los valores promedio en la especie, la tendencia se mantiene en términos generales en todos los conjuntos. Para determinar si los picos observados correspondían a distribuciones estadísticamente diferentes, se aplicó una prueba de bondad de ajuste de Kolmogorov-Smirnov, que es el equivalente de χ^2 para valores continuos.

Los resultados de esta prueba se muestran en una tabla en la misma fig.26 y, como puede verse, en todos los casos los valores de D son menores a 0.3 y los valores de P son mayores a 0.2, con lo cual podemos concluir que las distribuciones no son distintas entre sí.

Set	Num_genes	Mean_long	SD_long
set1	674	404.61	276.61
set2	574	104.15	50.95
set3	358	273.77	278.3
set4	1388	627.66	415.1
set5	923	544.34	477.1
set6	700	602	339.7
set7	1103	473.63	317.81
set8	605	359.96	288.69
set9	375	137.05	70.32
Global	6700	450.32	379.75

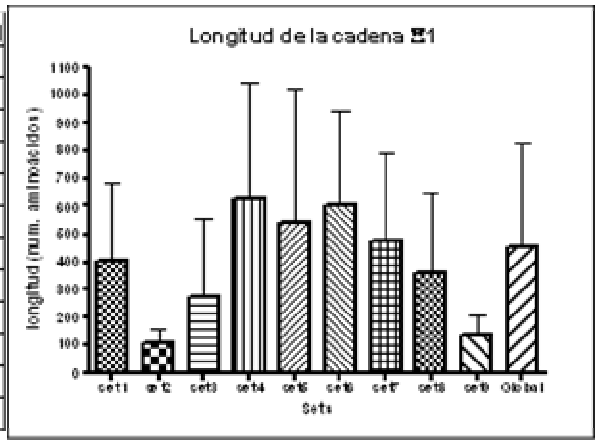


Tabla de valores de Z-score entre sets

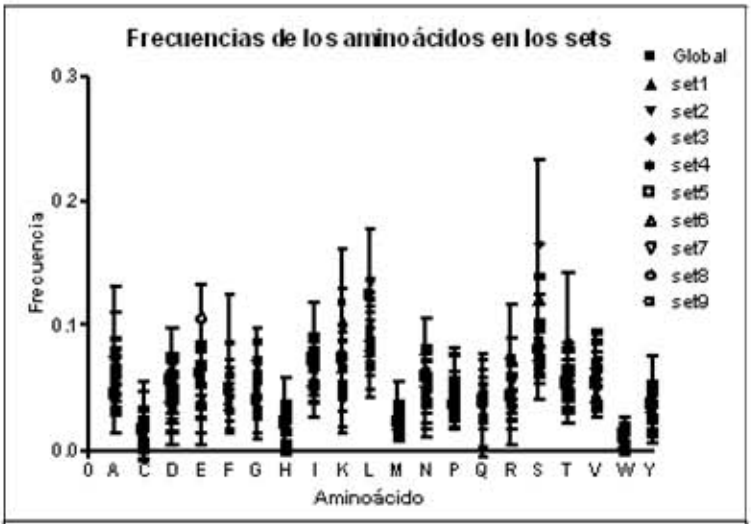
	set1	set2	set3	set4	set5	set6	set7	set8	set9	Global
set1	0	1.0862	0.473	-0.806	-0.505	-0.714	-0.25	0.1614	0.9673	-0.165
set2	**-5.897**	0	**-3.329**	**-10.27**	**-8.64**	**-9.771**	**-7.252**	**-5.021**	-0.646	**-6.794**
set3	-0.47	0.6095	0	-1.272	-0.972	-1.179	-0.718	-0.31	0.4913	-0.634
set4	0.5373	1.2612	0.8525	0	0.2007	0.0618	0.3711	0.6449	1.1819	0.4272
set5	0.2929	0.9226	0.5671	-0.175	0	-0.121	0.1482	0.3865	0.8537	0.1971
set6	0.5811	1.4656	0.9662	-0.076	0.1697	0	0.3779	0.7125	1.3687	0.4465
set7	0.2172	1.1626	0.6289	-0.485	-0.222	-0.404	0	0.3577	1.0591	0.0733
set8	-0.155	0.8861	0.2986	-0.927	-0.639	-0.838	-0.394	0	0.7721	-0.313
set9	**-3.805**	0.4679	-1.944	**-6.977**	**-5.792**	**-6.612**	**-4.786**	**-3.17**	0	**-4.455**
Global	0.1204	0.9116	0.4649	-0.467	-0.248	-0.399	-0.061	0.2379	0.8249	0

Fig. 25. Valores de longitud promedio de la cadena y *Z-score* para los conjuntos obtenidos a partir de Ξ_1.

La fig.27 muestra los resultados del análisis de longitud de la cadena para los conjuntos obtenidos a partir de Ξ_2. En este caso, podemos observar que los conjuntos 3, 7, 8 y 9 muestran diferencias con respecto a algunos de los demás conjuntos, pero no son significativamente diferentes entre sí y sólo el conjunto 3 y el 9 muestran diferencias mayores a 2 desviaciones estándar en relación a los valores globales de la especie. Estos datos nos permiten afirmar, que nuestra clasificación no se basa de manera exclusiva en la longitud de la cadena de aminoácidos.

Por su parte, la tabla A2.II muestra los valores obtenidos para la media y desviación estándar de la frecuencia de aminoácidos para cada uno de los conjuntos generados a partir de Ξ_2. En la fig.28 se muestran gráficamente estos resultados. Como puede notarse de la segunda gráfica, también en este caso se presentan algunos picos de distinta frecuencia para algunos aminoácidos en algunos conjuntos, pero la prueba de Kolmogorov-Smirnov para comparar

distribuciones, mostró que la distribución en la composición de aminoácidos no difiere significativamente entre conjuntos.



Prueba de bondad de ajuste de Kolmogorov-Smirnov

	set2		set3		set4		set5		set6		set7		set8		set9		Global	
	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P
set1	0.2	0.77	0.2	0.77	0.2	0.77	0.15	0.97	0.2	0.77	0.15	0.97	0.15	0.97	0.1	1	0.2	0.77
set2			0.15	0.97	0.2	0.77	0.2	0.77	0.2	0.77	0.25	0.5	0.2	0.77	0.2	0.77	0.25	0.5
set3					0.2	0.77	0.15	0.97	0.25	0.5	0.3	0.28	0.15	0.97	0.2	0.77	0.25	0.5
set4							0.15	0.97	0.1	1	0.15	0.97	0.1	1	0.15	0.97	0.15	0.97
set5									0.15	0.97	0.2	0.77	0.15	0.97	0.1	1	0.15	0.97
set6											0.25	0.5	0.1	1	0.15	0.97	0.15	0.97
set7													0.25	0.5	0.15	0.97	0.25	0.5
set8															0.15	0.97	0.15	0.97
set9																	0.15	0.97

Fig. 26. Comportamiento de la frecuencia promedio de cada aminoácido y resultados de la prueba de Kolmogorov-Smirnov entre todos los conjuntos obtenidos a partir de Ξ_1.

	Num_genes	Mean_long	SD_long
set1	1082	499.3	320.1
set2	722	457.5	282.6
set3	538	138	78.12
set4	1320	687.9	435.9
set5	1027	514	438.1
set6	599	607.4	343.1
set7	525	216.1	155.1
set8	558	231.7	167.3
set9	329	90.83	76.8
Global	6700	450.32	379.75

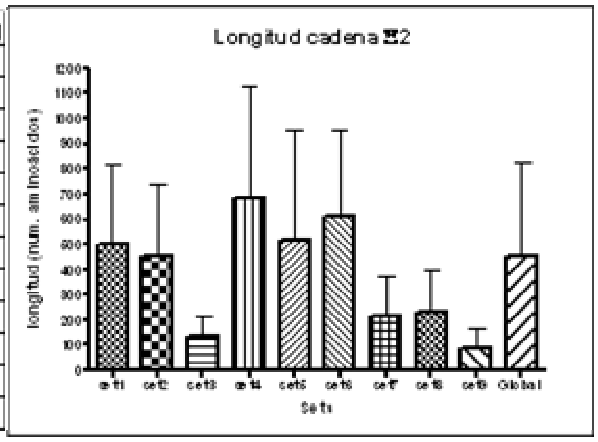


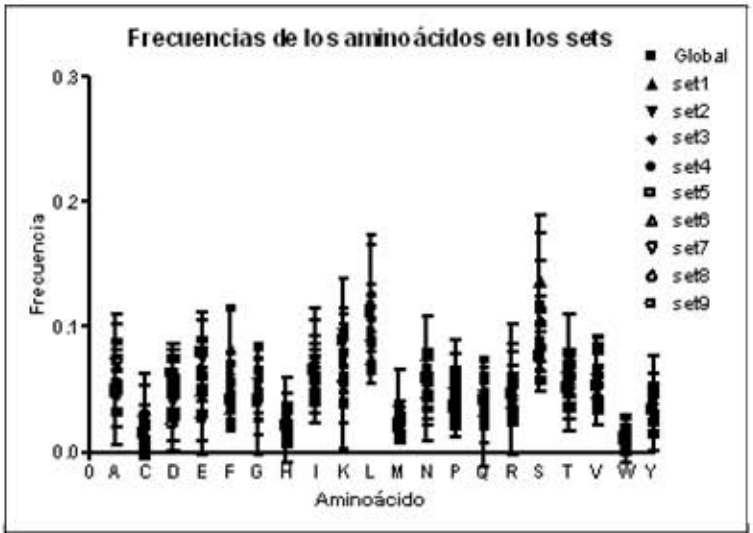
Tabla de valores de Z-score entre sets

	set1	set2	set3	set4	set5	set6	set7	set8	set9	Global
set1	0	0.1306	1.1287	-0.589	-0.046	-0.338	0.8847	0.836	1.2761	0.153
set2	-0.148	0	1.1306	-0.815	-0.2	-0.53	0.8542	0.799	1.2975	0.0254
set3	**-4.625**	**-4.09**	0	**-7.039**	**-4.813**	**-6.009**	-1	-1.199	0.6038	**-3.998**
set4	0.4327	0.5286	1.2615	0	0.3989	0.1847	1.0824	1.0466	1.3697	0.545
set5	0.0336	0.129	0.8583	-0.397	0	-0.213	0.68	0.6444	0.9659	0.1454
set6	0.3151	0.4369	1.3681	-0.235	0.2722	0	1.1405	1.095	1.5056	0.4578
set7	-1.826	-1.556	0.5035	**-3.042**	-1.921	**-2.523**	0	-0.101	0.8077	-1.51
set8	-1.6	-1.35	0.5601	**-2.727**	-1.687	**-2.246**	0.0932	0	0.842	-1.307
set9	**-5.319**	**-4.774**	-0.614	**-7.774**	**-5.51**	**-6.726**	-1.631	-1.834	0	**-4.681**
Global	0.129	0.0189	-0.822	0.6256	0.1677	0.4136	-0.617	-0.576	-0.947	0

Fig. 27. Valores de longitud promedio de la cadena y *Z-score* para los conjuntos obtenidos a partir de Ξ_2.

La fig.29 muestra el análisis de longitud de la cadena efectuado sobre los conjuntos obtenidos a partir de Ξ_3. En este caso, como en los dos anteriores, también hay algunos conjuntos que muestran diferencias con respecto a algunos otros. En particular, los conjuntos 4, 7, 8 y 9 muestran diferencias con algunos de los otros, pero ninguno de ellos presenta diferencias con todos los demás. Los conjuntos 7 y 9, que son los que se muestran más distintos con respecto a los otros, no muestran diferencias entre ellos con respecto a la longitud de la cadena de aminoácidos.

La tabla A2.III muestra los valores del promedio y la desviación estándar de la frecuencia para cada aminoácido en todos los conjuntos generados a partir de Ξ_3. La fig.30 muestra gráficamente estos datos. En la segunda gráfica se puede notar, como en los dos casos anteriores, que aunque algunos conjuntos (en particular el 7 y el 9) muestran fluctuaciones importantes respecto a la frecuencia promedio de la especie, la tendencia general es bastante uniforme en todos los conjuntos. La prueba de Kolmogorov-Smirnov aplicada para comparar las distribuciones de frecuencias entre conjuntos, mostró que ninguna de las distribuciones es distinta de las demás, y tampoco se observan diferencias con respecto a la distribución global de la especie.



Prueba de bondad de ajuste de Kolmogorv-Smirnov

	set2		set3		set4		set5		set6		set7		set8		set9		Global	
	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P
set1	0.2	0.77	0.25	0.5	0.15	0.97	0.15	0.97	0.25	0.5	0.2	0.77	0.25	0.5	0.25	0.5	0.25	0.5
set2			0.2	0.77	0.2	0.77	0.1	1	0.15	0.97	0.2	0.77	0.2	0.77	0.15	0.97	0.2	0.77
set3					0.25	0.5	0.15	0.97	0.2	0.77	0.35	0.14	0.25	0.5	0.15	0.97	0.25	0.5
set4							0.15	0.97	0.15	0.97	0.2	0.77	0.2	0.77	0.2	0.77	0.2	0.77
set5									0.1	1	0.25	0.5	0.2	0.77	0.15	0.97	0.15	0.97
set6											0.25	0.5	0.15	0.97	0.15	0.97	0.15	0.97
set7													0.1	1	0.2	0.77	0.15	0.97
set8															0.2	0.77	0.1	1
set9																	0.2	0.77

Fig. 28. Comportamiento de la frecuencia promedio de cada aminoácido y resultados de la prueba de Kolmogorov-Smirnov entre todos los conjuntos obtenidos a partir de Ξ_2.

	Num_genes	Mean_long	SD_long
set1	1116	487.8	312.6
set2	1446	673.2	427.1
set3	613	633.1	346.4
set4	593	214.3	155.5
set5	611	339.5	243
set6	1018	563.8	429.8
set7	222	69.89	48.96
set8	545	222.2	162.5
set9	536	123.2	56.32
Global	6700	450.32	379.75

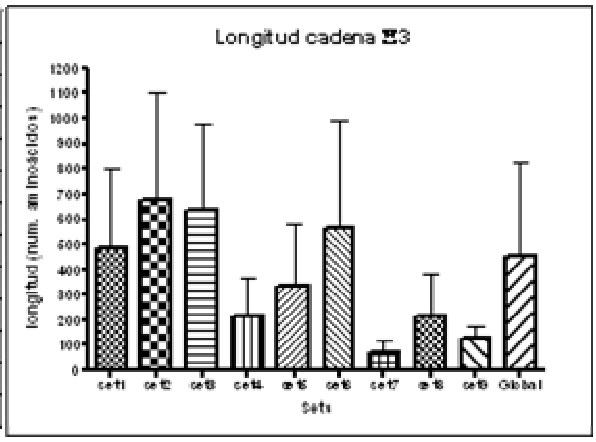


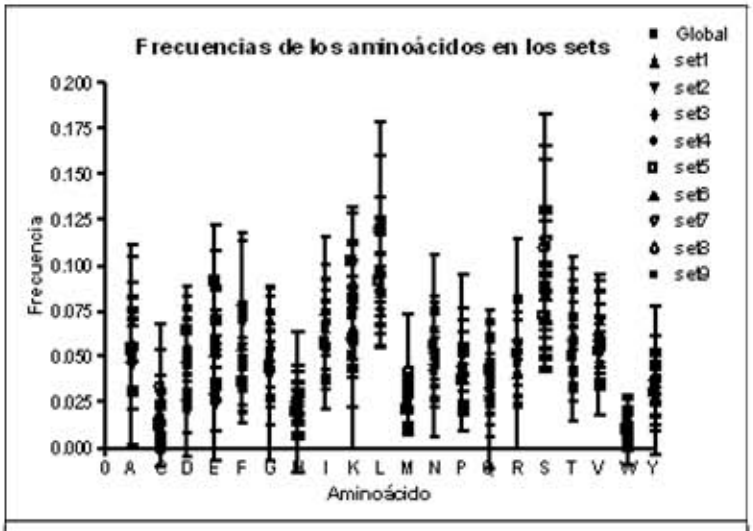
Tabla de valores de Z-score entre sets

	set1	set2	set3	set4	set5	set6	set7	set8	set9	Global
set1	0	-0.593	-0.465	0.8749	0.4744	-0.243	1.3369	0.8496	1.1663	0.1199
set2	0.4341	0	0.0939	1.0745	0.7813	0.2561	1.4126	1.056	1.2878	0.5218
set3	0.4195	-0.116	0	1.209	0.8476	0.2001	1.6259	1.1862	1.472	0.5277
set4	-1.759	**-2.951**	**-2.693**	0	-0.805	**-2.248**	0.9287	-0.051	0.5859	-1.518
set5	-0.61	-1.373	-1.208	0.5152	0	-0.923	1.1095	0.4827	0.8901	-0.456
set6	0.1768	-0.255	-0.161	0.8132	0.5219	0	1.1492	0.7948	1.0251	0.264
set7	**-8.536**	**-12.32**	**-11.5**	**-2.95**	**-5.507**	**-10.09**	0	**-3.111**	-1.089	**-7.77**
set8	-1.634	**-2.775**	**-2.529**	0.0486	-0.722	**-2.102**	0.9373	0	0.6092	-1.404
set9	**-6.474**	**-9.766**	**-9.054**	-1.618	**-3.841**	**-7.823**	0.9466	-1.758	0	**-5.808**
Global	0.0987	0.5869	0.4813	-0.622	-0.292	0.2988	-1.002	-0.601	-0.861	0

Fig. 29. Valores de longitud promedio de la cadena y *Z-score* para los conjuntos obtenidos a partir de Ξ_3.

Por último, se llevó a cabo el mismo análisis para los conjuntos obtenidos a partir de Ξ_4. La fig.31 resume los resultados obtenidos para el análisis de la longitud de las cadenas. Como en los casos anteriores, puede verse que algunos de los conjuntos presentan diferencias en cuanto a longitud promedio de sus cadenas de aminoácidos con algunos otros conjuntos, pero no hay ninguno que sea diferente de todos los demás.

La tabla A2.IV presenta los valores del promedio y desviación estándar de la frecuencia de aminoácidos en cada uno de los conjuntos obtenidos a partir de Ξ_4. La fig.32 muestra tanto las gráficas como el análisis de estos resultados. De la segunda gráfica puede verse que nuevamente se presentan picos en la distribución de frecuencias de algunos aminoácidos en algunos sets, pero la tendencia general no se desvía de manera significativa de la distribución de frecuencias global de la especie. Las comparaciones efectuadas mediante la prueba de Kolmogorov-Smirnov, tampoco muestran diferencias entre las distribuciones.



Prueba de bondad de ajuste de Kolmogorov-Smirnov

	set2		set3		set4		set5		set6		set7		set8		set9		Global	
	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P
set1	0.15	0.97	0.2	0.77	0.2	0.77	0.2	0.77	0.25	0.5	0.25	0.5	0.15	0.97	0.25	0.5	0.2	0.77
set2			0.15	0.97	0.15	0.97	0.1	1	0.2	0.77	0.2	0.77	0.2	0.77	0.2	0.77	0.2	0.77
set3					0.15	0.97	0.1	1	0.1	1	0.15	0.97	0.25	0.5	0.2	0.77	0.15	0.97
set4							0.15	0.97	0.15	0.97	0.2	0.77	0.15	0.97	0.25	0.5	0.1	1
set5									0.1	1	0.1	1	0.2	0.77	0.15	0.97	0.15	0.97
set6											0.15	0.97	0.15	0.97	0.15	0.97	0.2	0.77
set7													0.15	0.97	0.15	0.97	0.2	0.77
set8															0.3	0.28	0.15	0.97
set9																	0.2	0.77

Fig. 30. Comportamiento de la frecuencia promedio de cada aminoácido y resultados de la prueba de Kolmogorov-Smirnov entre todos los conjuntos obtenidos a partir de Ξ_3.

	Num_genes	Mean_long	SD_long
set1	527	126.4	45.43
set2	1791	691.3	457.7
set3	1553	544.8	317.7
set4	232	127.5	67.92
set5	558	642.9	331
set6	820	285.3	168
set7	144	58.48	39.79
set8	574	210.7	154.1
set9	501	228.9	169.3
Global	6700	450.32	379.75

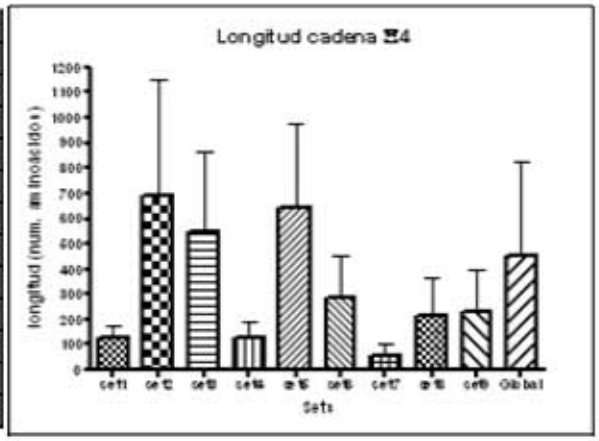


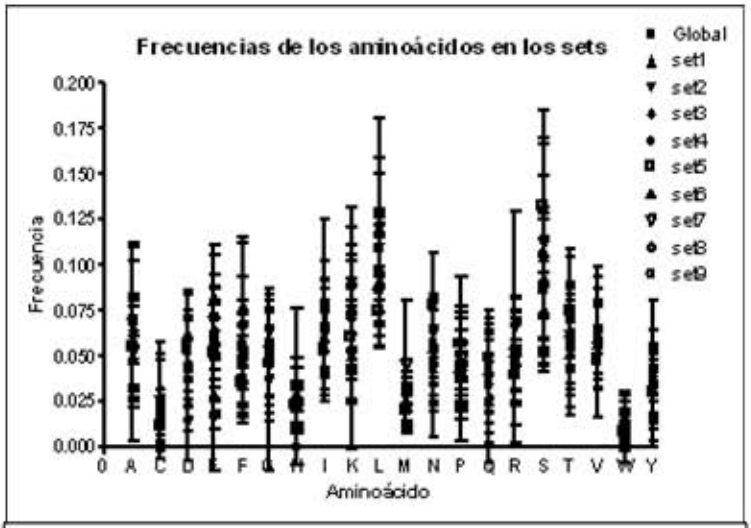
Tabla de valores de Z-score entre sets

	set1	set2	set3	set4	set5	set6	set7	set8	set9	Global
set1	0	**-12.435**	**-9.2098**	-0.0242	**-11.369**	**-3.4977**	1.49505	-1.8556	**-2.2562**	**-7.1301**
set2	1.23421	0	0.32008	1.23181	0.10575	0.88704	1.38261	1.05003	1.01027	0.5265
set3	1.31697	-0.4611	0	1.3135	-0.3088	0.81681	1.53075	1.05162	0.99433	0.29739
set4	0.0162	**-8.3009**	**-6.144**	0	**-7.5883**	**-2.3233**	1.0162	-1.225	-1.4929	**-4.7529**
set5	1.56042	-0.1462	0.29637	1.5571	0	1.08036	1.76562	1.30574	1.25076	0.58181
set6	0.94583	**-2.4167**	-1.5446	0.93929	**-2.1286**	0	1.35012	0.44405	0.33571	-0.9823
set7	-1.707	**-15.904**	**-12.222**	**-1.7346**	**-14.688**	**-5.7004**	0	**-3.8256**	**-4.283**	**-9.8477**
set8	0.54705	**-3.1188**	**-2.1681**	0.53991	**-2.8047**	-0.4841	0.9878	0	-0.1181	-1.555
set9	0.60543	**-2.7312**	-1.8659	0.59894	**-2.4454**	-0.3331	1.00662	0.1075	0	-1.3079
Global	-0.853	0.63458	0.2488	-0.8501	0.50712	-0.4345	-1.0318	-0.631	-0.5831	0

Fig. 31. Valores de longitud promedio de la cadena y *Z-score* para los conjuntos obtenidos a partir de Ξ_4.

Este análisis preliminar de los conjuntos obtenidos utilizando mapas de Kohonen nos permite afirmar que, aunque se presentan diferencias interesantes entre algunos conjuntos en términos de la longitud de la cadena y la composición de aminoácidos, éstos no son los únicos elementos que intervienen en la clasificación, ya que varios conjuntos presentan longitudes de cadena y distribuciones de frecuencia de aminoácidos semejantes, lo que nos permite afirmar que nuestro clasificador esta reflejando relaciones de orden superior entre los conjuntos de proteínas agrupados en los distintos conjuntos, evidentemente relacionadas con la estructura de las mismas.

Una vez descartada la hipótesis de que nuestro clasificador sea trivial, podemos proceder a hacer una validación de las clases obtenidas utilizando como referencia criterios funcionales, que nos permitirán determinar si los conjuntos se ajustan a categorías funcionales previamente descritas y por lo tanto reforzar la hipótesis original de que elementos estructurales de orden superior inmersos en la secuencia de aminoácidos presenten una estrecha relación con la funcionalidad de las proteínas.



Prueba de bondad de ajuste de Kolmogorov-Smirnov

	set2		set3		set4		set5		set6		set7		set8		set9		Global	
	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P	D	P
set1	0.15	0.97	0.25	0.5	0.15	0.97	0.2	0.77	0.2	0.77	0.2	0.77	0.25	0.5	0.25	0.5	0.2	0.77
set2			0.2	0.77	0.1	1	0.1	1	0.15	0.97	0.1	1	0.15	0.97	0.2	0.77	0.15	0.97
set3					0.2	0.77	0.25	0.5	0.2	0.77	0.2	0.77	0.2	0.77	0.15	0.97	0.2	0.77
set4							0.1	1	0.1	1	0.1	1	0.15	0.97	0.15	0.97	0.15	0.97
set5									0.15	0.97	0.15	0.97	0.15	0.97	0.2	0.77	0.15	0.97
set6											0.15	0.97	0.15	0.97	0.15	0.97	0.1	1
set7													0.2	0.77	0.25	0.5	0.15	0.97
set8															0.15	0.97	0.15	0.97
set9																	0.15	0.97

Fig. 32. Comportamiento de la frecuencia promedio de cada aminoácido y resultados de la prueba de Kolmogorov-Smirnov entre todos los conjuntos obtenidos a partir de Ξ_4.

4.3.2 Mapeo π (Validación de las clasificaciones)

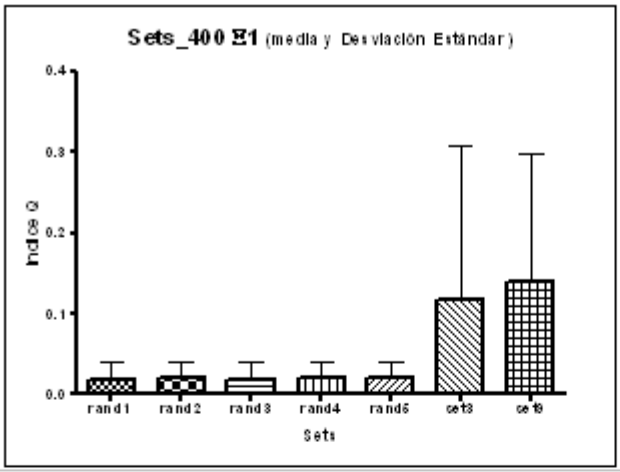
Como mencionamos en el inciso (3.4.1) de la metodología, utilizamos la base de datos Gene Ontology para validar nuestras clasificaciones. Según el procedimiento previamente descrito, calculamos los valores del índice Q para todos los conjuntos obtenidos a partir de las matrices Ξ_1-Ξ_4 así como para conjuntos aleatorios de proteínas de tamaños semejantes a nuestros conjuntos.

Nuevamente en esta sección, analizaremos los 9 conjuntos obtenidos a partir de cada una de nuestras matrices Ξ_1-Ξ_4. Para cada clasificación, analizaremos independientemente los conjuntos de tamaño semejante (en términos del número de genes que contienen), comparándolos con conjuntos aleatorios apropiados. Así, las figs.33-37 corresponden al análisis de los conjuntos obtenidos a partir de Ξ_1 (conjuntos de tamaño cercano a 400 secuencias fig.33, de tamaño cercano a 600 fig.34, de tamaño cercano a 1000 fig.35, de tamaño cercano a 1200 fig.36 y de tamaño cercano a 1400, fig.37), las figs.38-41, a los conjuntos obtenidos a partir de Ξ_2 (conjuntos de tamaño cercano a 400 fig.38, de tamaño cercano a 600 fig.39, de tamaño cercano a 1000 fig.40 y de tamaño cercano a 1400, fig.41), las figs.42-46 a los conjuntos obtenidos a partir de Ξ_3 (conjuntos de tamaño cercano a 200 fig.42, de tamaño cercano a 600 fig.43, de tamaño cercano a 1000 fig.44, de tamaño cercano a 1200 fig.45 y de tamaño cercano a 1400 fig.46) y por último, las figs.47-51 son las correspondientes al análisis del índice Q para los conjuntos obtenidos a partir de Ξ_4 (conjuntos de tamaño cercano a 200 fig.47, de tamaño cercano a 600 fig.48, de tamaño cercano a 800 fig.49, de tamaño cercano a 1600 fig.50 y de tamaño cercano a 1800 en la fig.51). En cada figura se presenta una gráfica que muestra el valor promedio y la desviación estándar del índice Q para cada conjunto, una tabla con dichos valores y las tablas con los resultados de los análisis estadísticos, correspondientes la primera a la comparación pareada entre nuestros conjuntos y cada conjunto aleatorio, utilizando la prueba de Mann Whitney y la segunda, a la prueba de Kruskal-Wallis resultante de comparar primero los conjuntos aleatorios entre sí y luego la totalidad de los conjuntos, aleatorios y no aleatorios.

En la fig.33 se presentan los resultados obtenidos para los conjuntos 3 y 9 de Ξ_1, de tamaño 358 y 375 respectivamente, comparados con conjuntos aleatorios de 400 genes. Como puede verse claramente en la gráfica, los valores del índice Q son mucho mejores para los conjuntos obtenidos mediante nuestro clasificador, lo que indica que se ajustan mejor a categorías funcionales de GO que los conjuntos aleatorios. En la primera tabla se presentan los resultados obtenidos mediante la prueba estadística de Mann Whitney, que permite hacer comparaciones pareadas entre las muestras. En todos los casos, los valores del índice Q de nuestros conjuntos fueron significativamente distintos ($P < 0.0001$) de los valores obtenidos para los conjuntos aleatorios. Por último, se presenta también una tabla con los resultados de la prueba estadística de Kruskal-Wallis entre varios conjuntos. Al aplicarla a los conjuntos aleatorios, se puede ver que éstos no difieren significativamente entre sí, pero al comparar la totalidad de los conjuntos, aleatorios y nuestros, sí hay diferencias significativas ($P<0.0001$) entre ellos.

Las fig.34 a la 37 presentan los mismos datos para los demás conjuntos de Ξ_1. En el caso de la fig.34 se presentan los resultados para los conjuntos 1 (de tamaño 674), 2 (574), 6 (700) y 8 (605), con respecto a conjuntos aleatorios de tamaño 600. Los resultados de los análisis estadísticos y las conclusiones a las que llegamos son las mismas que para el caso de los conjuntos de tamaño 400. En la fig. 35 se compara el set 5 (923) con conjuntos aleatorios de tamaño 1000, en la

fig.36 el set 7 (1103) con conjuntos de tamaño 1200, y por último, en la fig.37 el set 4 (1388) con conjuntos aleatorios de tamaño 1400. En todos los casos el análisis estadístico mostró que los conjuntos generados mediante nuestro clasificador se ajustan significativamente mejor a categorías funcionales de GO, que conjuntos aleatorios de tamaño semejante.



	Media Q	DS Q
rand1	0.01858	0.02184
rand2	0.01999	0.0208
rand3	0.01858	0.02184
rand4	0.01997	0.021
rand5	0.01999	0.0208
set3	0.1177	0.19
set9	0.1414	0.1563

Prueba de Mann Whitney

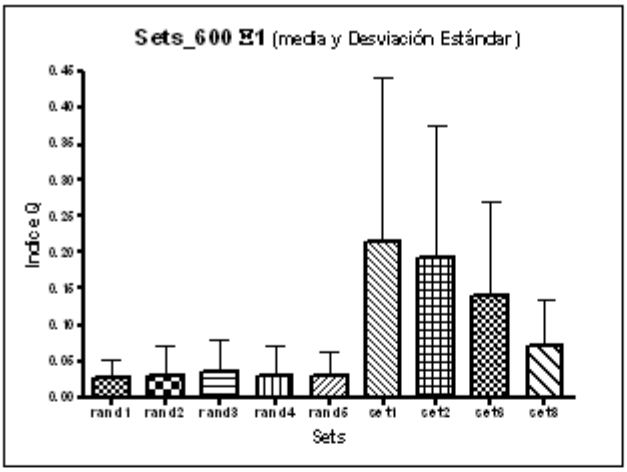
	set3_Vs_r1	set3_Vs_r2	set3_Vs_r3	set3_Vs_r4	set3_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
One- or two-tailed P value?	Two-tailed	Two-tailed	Two-tailed	Two-tailed	Two-tailed
Sum of ranks in columns	18996 , 19507	17661 , 18654	18996 , 19507	20478 , 19708	17661 , 18654
Mann-Whitney U	3596	3633	3596	4007	3633

	set9_Vs_r1	set9_Vs_r2	set9_Vs_r3	set9_Vs_r4	set9_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
One- or two-tailed P value?	Two-tailed	Two-tailed	Two-tailed	Two-tailed	Two-tailed
Sum of ranks in columns	18548 , 26903	17288 , 25783	18548 , 26903	20006 , 27272	17288 , 25783
Mann-Whitney U	3148	3260	3148	3535	3260

Prueba de Kruskal-Wallis

	random sets	all
P value	0.2081	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	7
Kruskal-Wallis statistic	5.883	233.3

Fig. 33. Valores del índice Q (media y desviación estándar) y resultados del análisis estadístico entre los conjuntos generados con nuestro clasificador y conjuntos aleatorios de tamaño cercano a 400 obtenidos a partir de Ξ_1.



	Media Q	DS Q
rand1	0.02675	0.02456
rand2	0.03049	0.04034
rand3	0.03434	0.04387
rand4	0.03049	0.04034
rand5	0.02965	0.03298
set1	0.2157	0.2269
set2	0.1941	0.1825
set6	0.1396	0.1316
set8	0.0707	0.06425

Prueba de Mann Whitney

	set1_Vs_r1	set1_Vs_r2	set1_Vs_r3	set1_Vs_r4	set1_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	5515	5609	6739	5609	5646

	set2_Vs_r1	set2_Vs_r2	set2_Vs_r3	set2_Vs_r4	set2_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	637	631	767	631	648

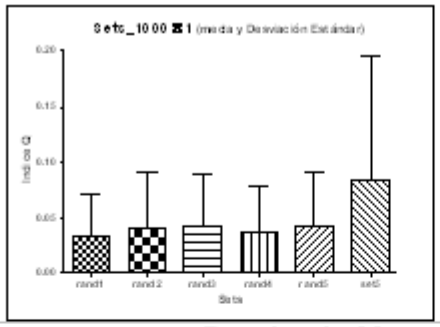
	set6_Vs_r1	set6_Vs_r2	set6_Vs_r3	set6_Vs_r4	set6_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	10940	11190	13400	11190	11420

	set8_Vs_r1	set8_Vs_r2	set8_Vs_r3	set8_Vs_r4	set8_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	16140	15950	19050	15950	16460

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.871	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	9
Kruskal-Wallis statistic	1.243	606.6

Fig. 34. Valores del índice Q (media y desviación estándar) y resultados del análisis estadístico entre los conjuntos generados con nuestro clasificador y conjuntos aleatorios de tamaño cercano a 600 obtenidos a partir de Ξ_1.



	Media Q	DS Q
rand1	0.0342	0.03719
rand2	0.04032	0.05057
rand3	0.04125	0.04811
rand4	0.03747	0.0408
rand5	0.04215	0.04981
set5	0.08409	0.1119

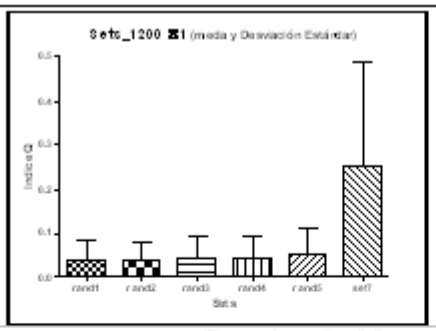
Prueba de Mann Whitney

	set5_Vs_r1	set5_Vs_r2	set5_Vs_r3	set5_Vs_r4	set5_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	36500	40740	42850	39880	39540

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.0757	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	8.475	116.3

Fig. 35. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1000 obtenidos a partir de Ξ_1.



	Media Q	DS Q
rand1	0.04275	0.04335
rand2	0.03942	0.0395
rand3	0.04385	0.04882
rand4	0.04447	0.04652
rand5	0.05265	0.05807
set7	0.2525	0.2377

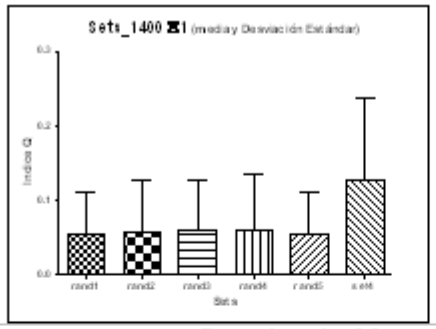
Prueba de Mann Whitney

	set7_Vs_r1	set7_Vs_r2	set7_Vs_r3	set7_Vs_r4	set7_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	33990	34200	34790	34610	39090

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.057	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	9.169	530.6

Fig. 36. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1200 obtenidos a partir de Ξ_1.



	Media Q	DS Q
rand1	0.05359	0.05809
rand2	0.05713	0.07037
rand3	0.06104	0.06486
rand4	0.06005	0.07391
rand5	0.05359	0.05809
set4	0.1261	0.1109

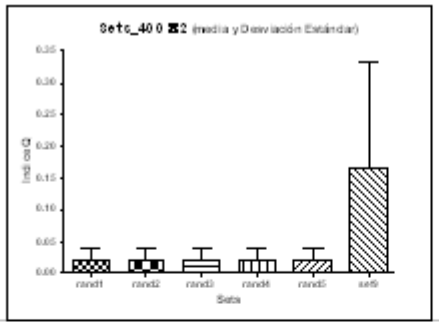
Prueba de Mann Whitney

	set4_Vs_r1	set4_Vs_r2	set4_Vs_r3	set4_Vs_r4	set4_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	50580	52520	57560	54750	50580

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.1172	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	7.379	232.8

Fig. 37. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1400 obtenidos a partir de Ξ_1.



	Media Q	DS Q
rand1	0.01858	0.02184
rand2	0.01999	0.0208
rand3	0.01858	0.02184
rand4	0.01997	0.021
rand5	0.01999	0.0208
set9	0.1645	0.1696

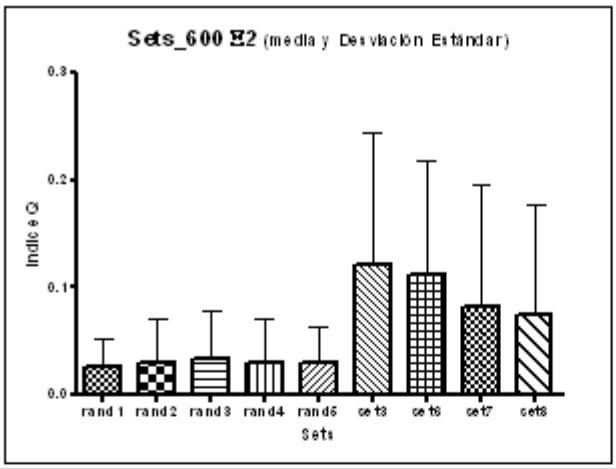
Prueba de Mann Whitney

	set9_Vs_r1	set9_Vs_r2	set9_Vs_r3	set9_Vs_r4	set9_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	422	488	422	495	488

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.2081	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	5.883	45.23

Fig. 38. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 400 obtenidos a partir de Ξ_2.



	Media Q	DS Q
rand1	0.02675	0.02456
rand2	0.03049	0.04034
rand3	0.03434	0.04387
rand4	0.03049	0.04034
rand5	0.02965	0.03298
set3	0.1208	0.1234
set6	0.1129	0.1057
set7	0.08206	0.1131
set8	0.07487	0.1016

Prueba de Mann Whitney

	set3_Vs_r1	set3_Vs_r2	set3_Vs_r3	set3_Vs_r4	set3_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	1394	1505	1874	1505	1551

	set6_Vs_r1	set6_Vs_r2	set6_Vs_r3	set6_Vs_r4	set6_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	10990	10960	13090	10960	11230

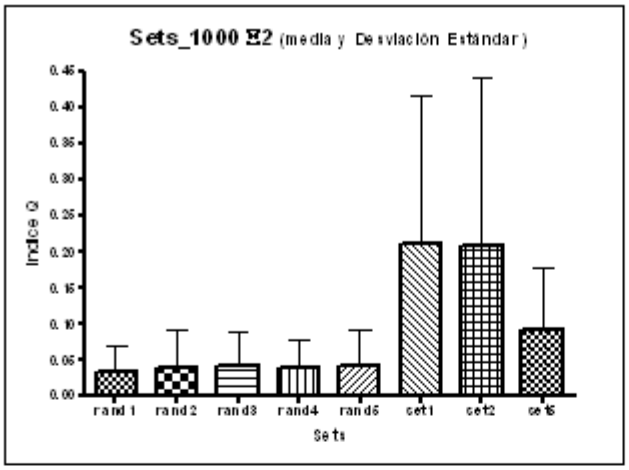
	set7_Vs_r1	set7_Vs_r2	set7_Vs_r3	set7_Vs_r4	set7_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	19250	18460	21200	18460	18980

	set8_Vs_r1	set8_Vs_r2	set8_Vs_r3	set8_Vs_r4	set8_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	23060	21820	25170	21820	22580

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.871	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	9
Kruskal-Wallis statistic	1.243	317.3

Fig. 39. Valores del índice Q (media y desviación estándar) y resultados del análisis estadístico entre los conjuntos generados con nuestro clasificador y conjuntos aleatorios de tamaño cercano a 600 obtenidos a partir de Ξ_2.



	Media Q	DS Q
rand1	0.0342	0.03719
rand2	0.04032	0.05057
rand3	0.04125	0.04811
rand4	0.03747	0.0408
rand5	0.04215	0.04981
set1	0.2113	0.2063
set2	0.2082	0.2331
set5	0.09119	0.08507

Prueba de Mann Whitney

	set1_Vs_r1	set1_Vs_r2	set1_Vs_r3	set1_Vs_r4	set1_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	26680	30800	32100	29600	29830

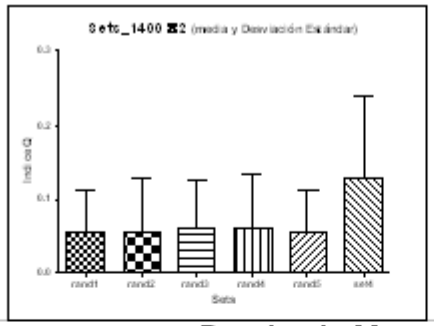
	set2_Vs_r1	set2_Vs_r2	set2_Vs_r3	set2_Vs_r4	set2_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	13480	15740	16490	15140	15190

	set5_Vs_r1	set5_Vs_r2	set5_Vs_r3	set5_Vs_r4	set5_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	29780	34330	35580	32850	32740

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.0757	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	8
Kruskal-Wallis statistic	8.475	705.7

Fig. 40. Valores del índice Q (media y desviación estándar) y resultados del análisis estadístico entre los conjuntos generados con nuestro clasificador y conjuntos aleatorios de tamaño cercano a 1000 obtenidos a partir de Ξ_2.



	Media Q	DS Q
rand1	0.05359	0.05809
rand2	0.05713	0.07037
rand3	0.06104	0.06486
rand4	0.06005	0.07391
rand5	0.05359	0.05809
set4	0.1295	0.1109

Prueba de Mann Whitney

	set4_Vs_r1	set4_Vs_r2	set4_Vs_r3	set4_Vs_r4	set4_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	49770	51100	56420	53760	49770

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.1172	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	7.379	217.4

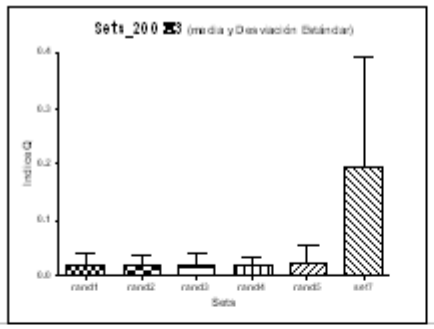
Fig. 41. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1400 obtenidos a partir de Ξ_2.

Las figs.38-41 muestran los resultados obtenidos para el análisis de los conjuntos obtenidos a partir de Ξ_2. En este caso, tenemos un conjunto de tamaño cercano a 400 (set9), cuatro de tamaño cercano a 600 (set3, 6, 7 y 8), dos conjuntos de tamaño cercano a 1000 y un conjunto de tamaño cercano a 1400 (set4). Los resultados del análisis de estos conjuntos son muy semejantes a los encontrados para el vector de una dimensión, es decir, el conjunto de valores del índice Q es distinto (P<0.0001) entre nuestros conjuntos y conjuntos aleatorios de tamaño semejante, lo que indica que nuestros conjuntos se ajustan mejor a categorías funcionales de GO que conjuntos de genes obtenidos de manera aleatoria.

En las figs.41-45 se presenta el resumen de los resultados obtenidos para el análisis de los conjuntos obtenidos a partir del vector de características de tres dimensiones. En este caso, se analizó un conjunto de tamaño cercano a 200 (set7), cinco conjuntos de tamaño cercano a 600 (set3, 4, 5, 8 y 9), un conjunto de tamaño cercano a 1000 (set6), un conjunto de tamaño cercano a 1200 (set1) y un conjunto de tamaño cercano a 1400 (set2). Aquí nuevamente, todos nuestros conjuntos mostraron diferencias significativas (P<0.0001) con respecto a los conjuntos aleatorios de tamaño semejante, por lo que podemos afirmar que nuestro clasificador genera conjuntos que se ajustan mejor a categorías funcionales de las proteínas que conjuntos aleatorios de genes.

Por último se analizaron los conjuntos obtenidos mediante nuestro clasificador a partir del vector de características de cuatro dimensiones. La fig.46 muestra el resultado del análisis de los conjuntos de tamaño cercano a 200 y mientras que el set7 muestra una enorme diferencia con respecto a los conjuntos aleatorios (P<0.0001), no ocurre lo mismo con el set4, cuya diferencia con tres de los conjuntos aleatorios, aunque significativa (P<0.005), no es tan grande como en

todos los conjuntos analizados hasta el momento. En la fig.47 se presentan los resultados del análisis de los conjuntos de tamaño cercano a 600 (set1, 5, 8 y 9). Aquí nuevamente tres conjuntos presentan diferencias muy grandes (P<0.0001) con los conjuntos aleatorios y uno de ellos (set9), presenta diferencias que aunque significativas, son menores que las anteriores (P<0.05). La fig.48 presenta el análisis del set6 de tamaño cercano a 800 que presenta diferencias muy significativas (P<0.0001) con respecto a los conjuntos aleatorios de tamaño semejante. Las figs.49 y 50 presentan el resultado del análisis de los sets 3 (de tamaño cercano a 1600) y 2 (de tamaño cercano a 1400). En ambos casos, los conjuntos obtenidos mediante nuestro clasificador son significativamente distintos (P<0.0001) de los conjuntos aleatorios de tamaño semejante, sin embargo en estos dos casos es muy importante hacer mención de que la prueba de Kruskal-Wallis entre los conjuntos aleatorios mostró diferencias significativas entre ellos. Este fenómeno se debe probablemente al hecho de que el tamaño de los conjuntos es muy grande e incluye cerca de una cuarta parte de todos los genes de la especie, lo cual puede resultar en que algunas categorías de GO presenten índices Q elevados debidos exclusivamente a la cantidad de genes analizados. Sin embargo, si observamos con detenimiento las gráficas en las que se representa la media y la desviación estándar del índice Q, puede notarse que las barras correspondientes a la media de Q para nuestros conjuntos rebasan considerablemente los límites de la combinación media+DS de todos los conjuntos aleatorios, por lo que cualquier prueba que se aplicara directamente sobre los valores promedio, como un Z-score reforzaría las diferencias encontradas y, por lo tanto, nuestra hipótesis de que los conjuntos obtenidos mediante nuestro clasificador presentan un mejor ajuste a categorías funcionales de GO, que conjuntos aleatorios de tamaño semejante.



	Media Q	DS Q
rand1	0.01929	0.01928
rand2	0.01828	0.01843
rand3	0.01801	0.02132
rand4	0.01724	0.01806
rand5	0.02349	0.03209
set7	0.1954	0.1974

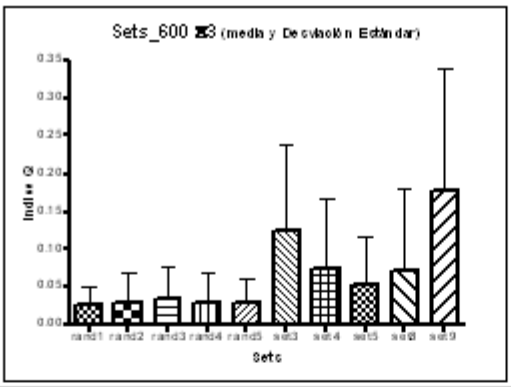
Prueba de Mann Whitney

	set7_Vs_r1	set7_Vs_r2	set7_Vs_r3	set7_Vs_r4	set7_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	203	164	152	195	233

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.764	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	1.846	31.63

Fig. 42. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 200 obtenidos a partir de Ξ_3.



	Media Q	DS Q
rand1	0.02675	0.02456
rand2	0.03049	0.04034
rand3	0.03434	0.04387
rand4	0.03049	0.04034
rand5	0.02965	0.03298
set3	0.1243	0.1156
set4	0.07362	0.09421
set5	0.05333	0.06303
set8	0.07136	0.1097
set9	0.1774	0.1629

Prueba de Mann Whitney

	set3_Vs_r1	set3_Vs_r2	set3_Vs_r3	set3_Vs_r4	set3_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	10200	10340	12430	10340	10580

	set4_Vs_r1	set4_Vs_r2	set4_Vs_r3	set4_Vs_r4	set4_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	20470	19500	22350	19500	20050

	set5_Vs_r1	set5_Vs_r2	set5_Vs_r3	set5_Vs_r4	set5_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	20880	20090	23460	20090	20850

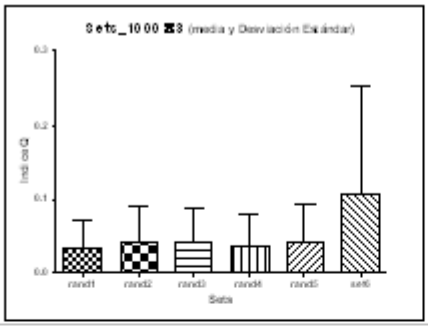
	set8_Vs_r1	set8_Vs_r2	set8_Vs_r3	set8_Vs_r4	set8_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	21370	20470	23390	20470	21050

	set9_Vs_r1	set9_Vs_r2	set9_Vs_r3	set9_Vs_r4	set9_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	282	360	470	360	353

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.871	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	10
Kruskal-Wallis statistic	1.243	354

Fig. 43. Valores del índice Q (media y desviación estándar) y resultados del análisis estadístico entre los conjuntos generados con nuestro clasificador y conjuntos aleatorios de tamaño cercano a 600 obtenidos a partir de Ξ_3.



	Media Q	DS Q
rand1	0.0342	0.03719
rand2	0.04032	0.05057
rand3	0.04125	0.04811
rand4	0.03747	0.0408
rand5	0.04215	0.04981
set6	0.1077	0.144

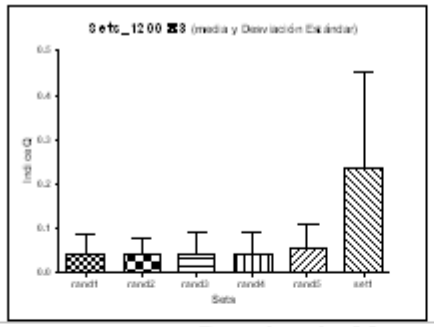
Prueba de Mann Whitney

	set6_Vs_r1	set6_Vs_r2	set6_Vs_r3	set6_Vs_r4	set6_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	34690	39140	40750	38060	37800

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.0757	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	8.475	157.7

Fig. 44. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1000 obtenidos a partir de Ξ_3.



	Media Q	DS Q
rand1	0.04275	0.04335
rand2	0.03942	0.0395
rand3	0.04385	0.04882
rand4	0.04447	0.04652
rand5	0.05265	0.05807
set1	0.2363	0.2193

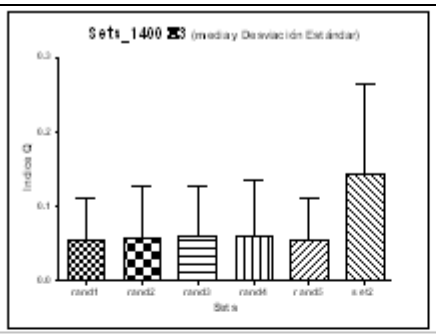
Prueba de Mann Whitney

	set1_Vs_r1	set1_Vs_r2	set1_Vs_r3	set1_Vs_r4	set1_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	29640	29840	30560	30500	34730

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.057	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	9.169	535.1

Fig. 45. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1200 obtenidos a partir de Ξ_3.



	Media Q	DS Q
rand1	0.05359	0.05809
rand2	0.05713	0.07037
rand3	0.06104	0.06486
rand4	0.06005	0.07391
rand5	0.05359	0.05809
set2	0.144	0.121

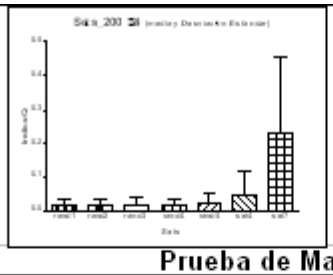
Prueba de Mann Whitney

	set2_Vs_r1	set2_Vs_r2	set2_Vs_r3	set2_Vs_r4	set2_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	43400	45390	49900	47800	43400

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.1172	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	7.379	319.4

Fig. 46. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1400 obtenidos a partir de Ξ_3.



	Media Q	DS Q
rand1	0.01929	0.01928
rand2	0.01828	0.01843
rand3	0.01801	0.02132
rand4	0.01724	0.01806
rand5	0.02349	0.03209
set4	0.0465	0.06792
set7	0.2291	0.2271

Prueba de Mann Whitney

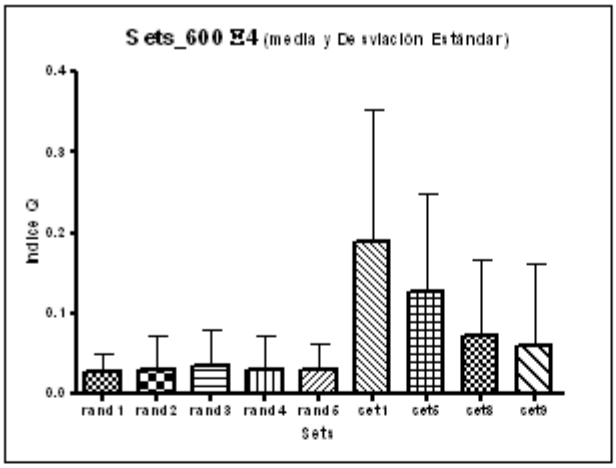
	set4_Vs_r1	set4_Vs_r2	set4_Vs_r3	set4_Vs_r4	set4_Vs_r5
P value	0.0003	0.0001	P<0.0001	P<0.0001	0.002
P value summary	***	***	***	***	**
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	2856	2504	2214	2927	3254

	set7_Vs_r1	set7_Vs_r2	set7_Vs_r3	set7_Vs_r4	set7_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	26	21	19	18	45

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.764	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	1.846	48.74

Fig. 47. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 200 obtenidos a partir de Ξ_4.



	Media Q	DS Q
rand1	0.02675	0.02456
rand2	0.03049	0.04034
rand3	0.03434	0.04387
rand4	0.03049	0.04034
rand5	0.02965	0.03298
set1	0.1884	0.1647
set5	0.1272	0.1231
set8	0.07328	0.09315
set9	0.06042	0.1019

Prueba de Mann Whitney

	set1_Vs_r1	set1_Vs_r2	set1_Vs_r3	set1_Vs_r4	set1_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	240	314	420	314	305

	set5_Vs_r1	set5_Vs_r2	set5_Vs_r3	set5_Vs_r4	set5_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	9490	9502	11490	9502	9734

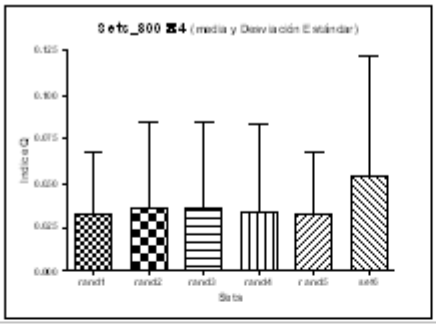
	set8_Vs_r1	set8_Vs_r2	set8_Vs_r3	set8_Vs_r4	set8_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	20950	20000	22780	20000	20480

	set9_Vs_r1	set9_Vs_r2	set9_Vs_r3	set9_Vs_r4	set9_Vs_r5
P value	0.0018	0.0014	0.0153	0.0014	0.0037
P value summary	**	**	*	**	**
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	21490	20360	23320	20360	21040

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.871	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	9
Kruskal-Wallis statistic	1.243	315.4

Fig. 48. Valores del índice Q (media y desviación estándar) y resultados del análisis estadístico entre los conjuntos generados con nuestro clasificador y conjuntos aleatorios de tamaño cercano a 600 obtenidos a partir de Ξ_4.



	Media Q	DS Q
rand1	0.03246	0.03511
rand2	0.03592	0.04859
rand3	0.0358	0.04888
rand4	0.03366	0.05034
rand5	0.03246	0.03511
set6	0.05395	0.06876

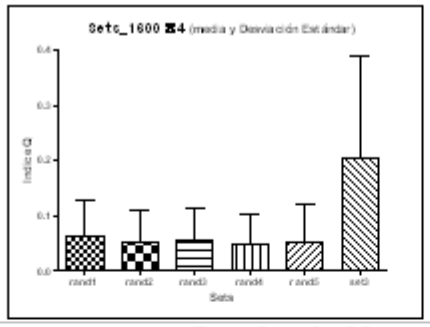
Prueba de Mann Whitney

	set6_Vs_r1	set6_Vs_r2	set6_Vs_r3	set6_Vs_r4	set6_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	34250	35500	36720	32500	34250

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.102	P<0.0001
P value summary	ns	***
Do the medians vary signif.	No	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	7.729	43.31

Fig. 49. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 800 obtenidos a partir de Ξ_4.



	Media Q	DS Q
rand1	0.0629	0.0648
rand2	0.05172	0.06111
rand3	0.05727	0.0558
rand4	0.05019	0.05375
rand5	0.05247	0.07009
set3	0.2062	0.1849

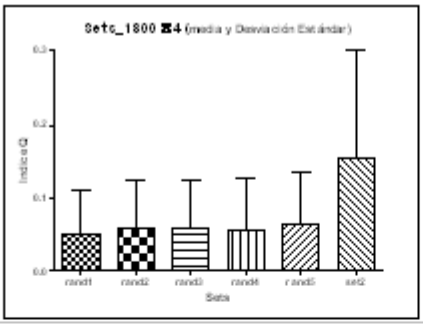
Prueba de Mann Whitney

	set3_Vs_r1	set3_Vs_r2	set3_Vs_r3	set3_Vs_r4	set3_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	61050	50930	53420	49430	47040

Prueba de Kruskal-Wallis

	random_sets	all
P value	P<0.0001	P<0.0001
P value summary	***	***
Do the medians vary signif.	Yes (P<0.05)	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	26.96	508

Fig. 50. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1600 obtenidos a partir de Ξ_4.



	Media Q	DS Q
rand1	0.05134	0.05862
rand2	0.06018	0.06336
rand3	0.05895	0.06503
rand4	0.05775	0.06997
rand5	0.06355	0.07259
set2	0.1537	0.1461

Prueba de Mann Whitney

	set2_Vs_r1	set2_Vs_r2	set2_Vs_r3	set2_Vs_r4	set2_Vs_r5
P value	P<0.0001	P<0.0001	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***	***	***
Are medians signif. different?	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)	Yes (P<0.05)
Mann-Whitney U	61040	69960	67910	61910	72270

Prueba de Kruskal-Wallis

	random_sets	all
P value	0.0111	P<0.0001
P value summary	*	***
Do the medians vary signif.	Yes (P<0.05)	Yes (P<0.05)
Number of groups	5	6
Kruskal-Wallis statistic	13.03	317.4

Fig. 51. Valores del índice Q (media y DS) y resultados del análisis estadístico entre los conjuntos de tamaño cercano a 1800 obtenidos a partir de Ξ_4.

De los análisis realizados hasta el momento, podemos concluir que los conjuntos obtenidos mediante nuestro clasificador utilizando los vectores de características Ξ_1-Ξ_4, muestran una correlación interesante con categorías funcionales de las proteínas, según han sido descritas en la base de datos Gene Ontology. Aunque estos resultados no nos permiten discriminar entre las distintas formas de representar las secuencias (Ξ_1-Ξ_4) en términos del ajuste de los conjuntos generados a partir de cada una de ellas a categorías funcionales de GO, si podemos anticipar que los datos obtenidos para Ξ_4 parecen ser los menos robustos de las cuatro representaciones. Los análisis realizados sobre el índice Q para todos los conjuntos nos indican que nuestro algoritmo de clasificación de secuencias de proteínas nos está permitiendo generar conjuntos que presentan relaciones funcionales interesantes entre ellos, identificables como términos de GO bien representados dentro de cada conjunto.

Es muy importante mencionar que el ajuste que estamos observando, no responde únicamente al hecho de que nodos aislados del DAG de GO se encuentran muy bien representados en nuestros conjuntos, sino que cuando observamos la posición de las categorías más significativas para cada conjunto a nivel del árbol de GO, puede notarse que son ramas completas del árbol las que se muestran pobladas por cada uno de los conjuntos y generalmente distintas ramas para los distintos conjuntos.

El diagrama de la fig.52 trata de ejemplificar el concepto antes descrito. En el primer DAG (**a**) puede observarse que hay 8 categorías significativas (marcadas en rojo), pero éstas se encuentran dispersas en el árbol y no forman parte de ramas distinguibles. Por el contrario, en el segundo DAG (**b**) puede notarse que las categorías significativas están agrupadas entre sí, ocupando ramas completas del diagrama, como ocurre con la mayoría de nuestros conjuntos. Dado que las ramas están ordenadas de acuerdo con un criterio jerárquico y presentan una relación entre ellas, este hecho es importante, ya que indica que cada uno de nuestros conjuntos contiene genes asociados con alguna función específica, pero que al mismo tiempo comparten una función más general entre sí.

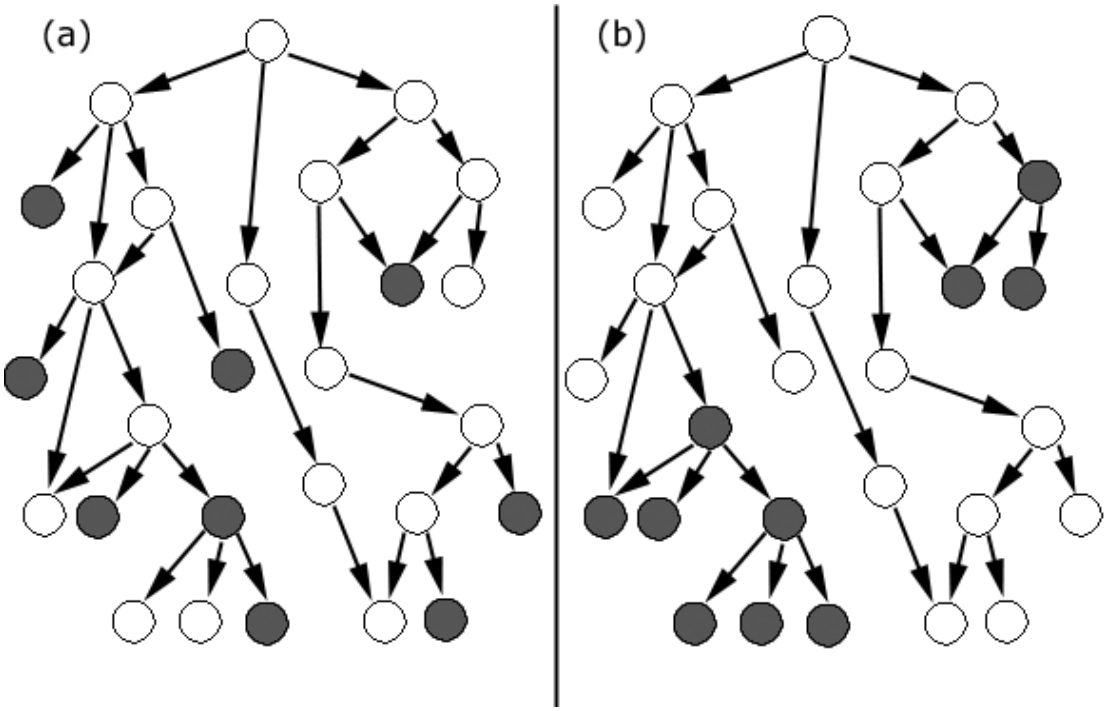


Fig. 52. Muestra el esquema de dos DAG´s. En el primero (a) puede notarse que los nodos significativos (oscuros) se encuentran dispersos, mientras que en el segundo (b) los nodos significativos se agrupan ocupando ramas completas de la gráfica.

Con el fin de ilustrar de manera muy general este concepto de ajuste a categorías funcionales a continuación presentaré algunos datos que muestran cuáles son las categorías funcionales más significativas para algunos de los conjuntos obtenidos a partir de Ξ_1 así como un resumen de las categorías más significativas en todos los demás conjuntos obtenidos a partir de Ξ_2- Ξ_4.[9]

La fig.53 muestra los términos de GO más significativos para los sets 1 y 6 de la clasificación obtenida a partir de Ξ_1. Como puede notarse, el conjunto 1 está claramente dominado por proteínas relacionadas con la función de transporte, de hecho, 25 de los primeros 30 términos con un mayor índice Q en este conjunto

[9] A partir de este momento, siempre que incluya la descripción de un término de GO, lo haré utilizando la nomenclatura original, en inglés, para evitar confusiones por diferencias en la traducción.

corresponden a procesos biológicos asociados con el transporte. Por su parte, el set6 está representado fundamentalmente por proteínas reguladoras. Para la figura se tomaron exclusivamente los primeros 17 términos de GO que presentaron altos valores para el índice Q y de éstos, 13 están relacionados con algún tipo de regulación. En la fig.54 se presentan los mismos datos para el set7, que muestra una mayor representatividad de proteínas asociadas con el metabolismo celular. Patrones semejantes aunque menos obvios pueden encontrarse en casi todos los conjuntos generados para las cuatro matrices utilizadas.

Términos de GO más representativos del set1

GO:0015892	iron-siderophore **transport**	GO:0046942	carboxylic acid **transport**
GO:0000101	sulfur amino acid **transport**	GO:0015849	organic acid **transport**
GO:0015833	peptide **transport**	GO:0015846	polyamine **transport**
GO:0006672	ceramide metabolism	GO:0015891	siderophore **transport**
GO:0006828	manganese ion **transport**	GO:0006874	calcium ion homeostasis
GO:0015802	basic amino acid **transport**	GO:0006829	zinc ion **transport**
GO:0006865	amino acid **transport**	GO:0006817	phosphate **transport**
GO:0008645	hexose **transport**	GO:0015893	drug **transport**
GO:0015749	monosaccharide **transport**	GO:0006665	sphingolipid metabolism
GO:0015837	amine **transport**	GO:0030148	sphingolipid biosynthesis
GO:0015875	vitamin or cofactor **transport**	GO:0015718	monocarboxylic acid **transport**
GO:0015804	neutral amino acid **transport**	GO:0000041	transition metal ion **transport**
GO:0006882	zinc ion homeostasis	GO:0006820	anion **transport**
GO:0015677	copper ion **import**	GO:0015698	inorganic anion **transport**
GO:0008643	carbohydrate **transport**	GO:0015711	organic anion **transport**

Términos de GO más representativos del set6

GO:0050793	**regulation** of development
GO:0040008	**regulation** of growth
GO:0051049	**regulation** of transport
GO:0006110	**regulation** of glycolysis
GO:0008361	**regulation** of cell size
GO:0006808	**regulation** of nitrogen utilization
GO:0019933	cAMP-mediated signaling
GO:0019935	cyclic-nucleotide-mediated signaling
GO:0045815	positive **regulation** of gene expression, epigenetic
GO:0006345	loss of chromatin silencing
GO:0045893	positive **regulation** of transcription, DNA-dependent
GO:0007124	pseudohyphal growth
GO:0045944	positive **regulation** of transcription from RNA polymerase II promoter
GO:0009894	**regulation** of catabolism
GO:0009893	positive **regulation** of metabolism
GO:0045935	positive **regulation** of nucleobase, nucleoside, nucleotide and...
GO:0045941	positive **regulation** of transcription

Fig. 53. Términos de GO asociados a los valores más altos del índice Q para las proteínas agrupadas en los conjuntos 1 y 6 de la clasificación generada a partir de Ξ_1.

Términos de GO más representativos del set7

GO:0009070	serine family amino acid biosynthesis	GO:0019320	hexose **catabolism**
GO:0006098	pentose-phosphate shunt	GO:0006007	glucose **catabolism**
GO:0042816	vitamin B6 **metabolism**	GO:0046040	IMP **metabolism**
GO:0009123	nucleoside monophos. **metabolism**	GO:0006551	leucine **metabolism**
GO:0006730	one-carbon compound **metabolism**	GO:0009228	thiamin biosynthesis
GO:0006541	glutamine **metabolism**	GO:0008614	pyridoxine **metabolism**
GO:0006542	glutamine biosynthesis	GO:0006595	polyamine **metabolism**
GO:0006544	glycine **metabolism**	GO:0006739	NADP **metabolism**
GO:0042219	amino acid derivative **catabolism**	GO:0000255	allantoin **metabolism**
GO:0009069	serine family amino acid **metabolism**	GO:0000256	allantoin **catabolism**
GO:0019794	nonprotein amino acid **metabolism**	GO:0019660	glycolytic fermentation
GO:0006566	threonine **metabolism**	GO:0006591	ornithine **metabolism**

Fig. 54. Términos de GO asociados a los valores más altos del índice Q para las proteínas agrupadas en el conjunto 7 de la clasificación generada a partir de Ξ_1.

En las figs.55 y 56 se presenta un listado de las grandes categorías representativas de los genes de cada uno de los conjuntos generados mediante nuestro clasificador, es decir raíces de subárboles que están altamente pobladas por nodos que resultaron ser significativos en cada uno de los conjuntos analizados. También en estas figuras se incluye una columna que indica el porcentaje de genes de función desconocida (aún no anotados) asociados a cada conjunto.

En este punto es interesante notar que los genes de función desconocida no están distribuidos de manera uniforme en nuestros conjuntos. Mientras que el porcentaje de genes no anotados para la especie es de 33.7%, nuestros conjuntos presentan valores que son o mucho mayores o ligeramente menores que éste. El hecho de que al menos dos conjuntos de cada clasificación estén dominados por proteínas de función desconocida (en varios casos con porcentajes cercanos al 90%), parece indicar que una fracción importante de los genes no anotados no estén asociados a procesos biológicos conocidos, sino probablemente correspondan a procesos muy específicos aún no estudiados y en los cuales participan proteínas con estructuras notablemente distintas a las ya descritas para los procesos generales de transporte, metabolismo y ciclo celular. Por otro lado, en los conjuntos donde se agrupan gran cantidad de genes asociados con las funciones más generales, los porcentajes de genes de función desconocida tienden a ser menores al 33% de la especie, lo que indicaría que estos procesos biológicos son los más conocidos y estudiados y por lo mismo, son pocos los genes asociados a ellos que aún no han sido descritos.

Conjuntos obtenidos a partir de Ξ1		
set	Categorías funcionales de GO	Función desconocida
1	Transporte, Homeostasis de cationes	42%
2	Biosíntesis y metabolismo de Glycoproteínas	93.20%
3	Adhesión celular, respuesta al calor, respuesta a feromonas	64%
4	Ciclo celular, meiosis, mitosis, organización y biogénesis de cromosomas, metabolismo de DNA	25.40%
5	Biosíntesis y metabolismo de lipoproteínas, transporte mediado por vesículas	31.50%
6	Regulación, regulación del metabolismo de ácidos nucléicos, importación de proteínas al núcleo, recombinación de DNA	24.40%
7	Metabolismo y biosíntesis celular	20.60%
8	Biogénesis y ensamble de ribosomas, organización y biogénesis de organelos, metabolismo de RNA ribosomal	25%
9	Metabolismo y biosíntesis de nucleótidos, metabolismo de ATP, transporte de hidrógeno, respuesta a estrés hídrico	32%

Conjuntos obtenidos a partir de Ξ2		
set	Categorías funcionales de GO	Función desconocida
1	Metabolismo y biosíntesis celular, metabolismo de aminoácidos, metabolismo de carbohidratos	21.30%
2	Transporte, homeostasis de cationes	33.20%
3	Metabolismo de fosfolípidos	87%
4	Regulación del metabolismo de DNA, recombinación de DNA, replicación de DNA	24.50%
5	Ciclo celular, organización y biogénesis de organelos, metabolismo de RNA, segregación de cromosomas	25%
6	Adhesión celular, biosíntesis de polisacáridos, regulación de la transcripción	26.50%
7	Biosíntesis de ATP, metabolismo y biosíntesis de nucleótidos, respuesta a la desecación, transporte de hidrógeno, transporte de electrones	29%
8	Organización de la membrana mitocondrial, metabolismo y biosíntesis de nucleótidos	41.80%
9	Respuesta a substancias químicas	86.60%

Fig. 55. Categorías funcionales de GO que están representadas por varios términos significativos en cada uno de los conjuntos generados a partir de Ξ_1 y Ξ_2.

Conjuntos obtenidos a partir de Ξ3		
set	Categorías funcionales de GO	Función desconocida
1	Metabolismo y biosíntesis celular, metabolismo de derivados de aminoácidos, metabolismo de vitaminas	21.60%
2	Ciclo celular, mitosis, modificación de la cromatina, metabolismo, recombinación, reparación y replicación de DNA	23.40%
3	Desarrollo, Ciclo celular, Organización del citoplasma, morfogénesis, crecimiento	24.60%
4	Biosíntesis de proteínas, metabolismo y biosíntesis de nucleótidos, transcricpión, metabolismo de ATP, transporte de hidrógeno	48.90%
5	Biogénesis y ensamble de ribosomas, metabolismo y procesamiento de RNA ribosomal, replicación	28%
6	Transporte, Metabolismo y biosíntesis de lípidos, metabolismo de lípidos de membrana, homeostasis de cationes	30.70%
7	Respuesta a sustancias químicas, transporte de iones	89.20%
8	Metabolismo y biosíntesis de nucleótidos, transporte de electrones, respuesta al calor y respuesta a la desecación	30.30%
9	Transporte mediado por vesículas	91.20%

Conjuntos obtenidos a partir de Ξ4		
set	Categorías funcionales de GO	Función desconocida
1	Transporte mediado por vesículas	91.30%
2	Ciclo celular, mitosis, meiosis, metabolismo y biosíntesis de lípidos de membrana, lipidación de proteínas	25.70%
3	Metabolismo, catabolismo y respiración celular, envejecimiento, metabolismo de aminoácidos, biosíntesis de nucleótidos	20.50%
4	Organización y biogénesis de vacuola	63.80%
5	Adhesión celular, morfogénesis celular, crecimiento, metabolismo de polisacáridos celulares	24.70%
6	Biogénesis y ensamble de ribosomas, metabolismo de RNA ribosomal, procesamiento de RNA	28.90%
7	Respuesta a estímulos	91%
8	Metabolismo y biosíntesis de nucleótidos, metabolismo de nucleobases, metabolismo y biosíntesis de ATP, biosíntesis de compuestos aromáticos.	49.30%
9	Metabolismo y biosíntesis de nucleósidos trifosfato, fosforilación oxidativa, transporte de electrones, respuesta a la desecación	32.10%

Fig. 56. Categorías funcionales de GO que están representadas por varios términos significativos en cada uno de los conjuntos generados a partir de Ξ_3 y Ξ_4.

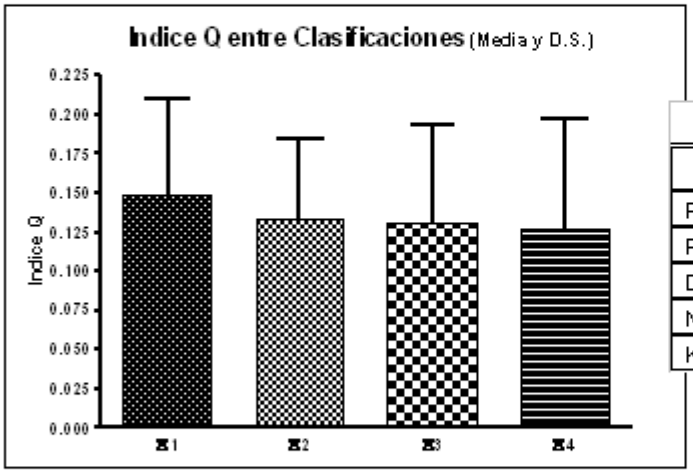
4.3.3 Comparación de las clasificaciones

Como último elemento de análisis de los resultados obtenidos hicimos una comparación entre las cuatro clasificaciones (Ξ_1-Ξ_4) en términos de los valores calculados para el índice Q para todos sus conjuntos. La finalidad de esta comparación es intentar discriminar entre los cuatro métodos de representación de las secuencias, para poder establecer si alguno de ellos es superior a los demás, en términos de la correlación con categorías funcionales de los conjuntos obtenidos.

La fig.57 muestra los resultados de este análisis de forma abreviada. En la primera tabla se muestran los valores de la media y la desviación estándar del índice Q para cada uno de los conjuntos obtenidos de las cuatro clasificaciones. Estos datos se presentan también en forma gráfica en la misma figura. Para determinar si las pequeñas diferencias que se observaban entre las clasificaciones tenían un significado estadístico, decidimos aplicar una prueba de Kruskal-Wallis entre las cuatro, misma que nos indicó que no existen diferencias significativas entre ellas. Para darle mayor fuerza a este resultado, decidimos hacer también comparaciones pareadas entre ellas. Los resultados de las comparaciones utilizando la U de Mann Whitney se presentan en la última tabla de la misma figura y como era de esperarse, los resultados son consistentes con lo obtenido mediante la prueba de Kruskal-Wallis, ninguna de las clasificaciones mostró diferencias significativas con todas las demás.

Estas comparaciones no proporcionaron suficiente información para poder discriminar entre las cuatro clasificaciones, por lo que hasta el momento no es posible establecer una de las cuatro representaciones de las secuencias como mejor que las otras en términos del análisis mediante nuestro estimador Q. Sería necesario diseñar algún otro procedimiento que nos permitiera hacer un análisis más fino de los conjuntos obtenidos y de esta forma poder escoger una representación de las secuencias como la más apropiada para la clasificación.

	Clasificación Ξ1		Clasificación Ξ2		Clasificación Ξ3		Clasificación Ξ4	
	Q Media	Q DS	Q Media	Q DS	Q Media	Q DS	Q Media	Q DS
set1	0.2157	0.2269	0.2113	0.2063	0.2363	0.2193	0.1884	0.1647
set2	0.1941	0.1825	0.2082	0.2331	0.144	0.121	0.1537	0.1461
set3	0.1177	0.19	0.1208	0.1234	0.1243	0.1156	0.2062	0.1849
set4	0.1261	0.1109	0.1295	0.1109	0.07362	0.09421	0.0465	0.06792
set5	0.08409	0.1119	0.09119	0.08507	0.05333	0.06303	0.1272	0.1231
set6	0.1396	0.1316	0.1129	0.1057	0.1077	0.144	0.05395	0.06876
set7	0.2525	0.2377	0.08206	0.1131	0.1954	0.1974	0.2291	0.2271
set8	0.0707	0.06425	0.07487	0.1016	0.07136	0.1097	0.07328	0.09315
set9	0.1414	0.1563	0.1645	0.1696	0.1774	0.1629	0.06042	0.1019



Prueba de Kruskal-Wallis	
	all
P value	0.8312
P value summary	ns
Do the medians vary signif.	No
Number of groups	4
Kruskal-Wallis statistic	0.8759

Prueba de Mann Whitney						
	Ξ1_Vs_Ξ2	Ξ1_Vs_Ξ3	Ξ1_Vs_Ξ4	Ξ2_Vs_Ξ3	Ξ2_Vs_Ξ4	Ξ3_Vs_Ξ4
P value	0.5457	0.6048	0.4894	0.7962	0.6048	0.8633
P value summary	ns	ns	ns	ns	ns	ns
Are medians signif. different?	No	No	No	No	No	No
One- or two-tailed P value?	Two-tailed	Two-tailed	Two-tailed	Two-tailed	Two-tailed	Two-tailed
Sum of ranks in columns	93 , 78	92 , 79	94 , 77	89 , 82	92 , 79	88 , 83
Mann-Whitney U	33	34	32	37	34	38

Fig. 57. Tabla y gráfica de valores del índice Q para los 9 conjuntos de cada una de las cuatro clasificaciones (Ξ_1-Ξ_4) y resumen de los análisis estadísticos que se utilizaron para compararlas.

4.4 Exploración de las potencialidades del método como clasificador de nuevas proteínas.

Habiendo demostrado que el método de clasificación propuesto genera clases de proteínas que se asocian de manera clara con categorías funcionales de GO, el siguiente paso sería explorar sus potencialidades para la clasificación de nuevos conjuntos de proteínas. Con esta finalidad, escogimos una de nuestras clasificaciones e hicimos una descripción funcional lo más precisa posible de cada una de sus clases. Luego, utilizando la red de Kohonen previamente entrenada, le proporcionamos como entrada un conjunto aleatorio de proteínas y analizamos el ajuste de los subconjuntos así obtenidos a las descripciones funcionales de cada una de las clases previamente descritas.

Dado que nuestro método de validación no nos permite escoger una de las cuatro clasificaciones (Ξ_1, Ξ_2, Ξ_3 o Ξ_4) como óptima (4.3.3), la elección de la clasificación a utilizar para este paso se hizo bajo criterios un poco subjetivos. Generamos los DAG´s de GO para cada una de las clases de las distintas clasificaciones y decidimos utilizar aquélla en la que con el menor número de descriptores pudiera abarcarse un mayor número de proteínas de los conjuntos (es decir, un mayor número de nodos agrupados en el menor número de ramas, como se mostró en la fig.53 [b]). De manera interesante, aunque la clasificación obtenida a partir de Ξ_1 mostró valores un poco mejores de ajuste a categorías funcionales según los datos de nuestro índice ($\overline{Q} \approx 0.15$), fue la clasificación obtenida mediante Ξ_3 ($\overline{Q} \approx 0.13$) la que mostró un mejor agrupamiento de las categorías de GO en ramas completas del árbol, permitiendo ésto una descripción más compacta de los conjuntos. Este dato es importante para nosotros, ya que parte de nuestra hipótesis original consistía en la suposición de que el ordenamiento de las secuencias en un mapa tridimensional podría proporcionar información interesante respecto a su estructura y en ese sentido, el hecho de que la matriz Ξ_3 haya mostrado un patrón de ajuste interesante a ramas completas del DAG de GO parece apoyar esta hipótesis, aunque sería necesario diseñar alguna herramienta que nos permitiera cuantificar estos patrones para poderlos contrastar entre clasificaciones.

En la tabla A3.I, correspondiente al tercer apéndice, se incluyen las descripciones completas para cada uno de los 9 conjuntos de proteínas obtenidos utilizando Ξ_3. Cabe señalar que los porcentajes asociados a cada término de GO se calcularon sin tomar en cuenta las proteínas desconocidas de cada clase (GO:0000004).

Para probar nuestro clasificador generamos 3 conjuntos aleatorios de genes del genoma de *Saccharomyces* de tamaño 500 y 3 conjuntos aleatorios de tamaño 1000. Cada uno de estos conjuntos, representado en la forma de vector de características, se presentó como entrada a la red de Kohonen previamente entrenada (4.2.2). Como resultado de la clasificación, las proteínas de cada una de las 6 muestras fueron separadas en las 9 clases (sets) previamente definidas. Por último, restaba comparar los 9 conjuntos obtenidos para cada una de nuestras muestras aleatorias recién clasificadas con las definiciones de las clases (A2.I), para determinar si los genes de cada clase se encontraban definidos de manera adecuada.

Con esta finalidad, desarrollamos una herramienta en perl (*ajuste_definicion.pl*) que a partir de un listado de términos de GO (definición de cada clase) y un conjunto de genes de entrada, va recorriendo el DAG de la

ontología y determina para cada uno de los genes si su función se encuentra contenida en el listado que representa la definición de la clase. Como resultado, el programa nos arroja dos estimadores: **i**) el porcentaje de genes de la muestra que está definido en el listado (porcentaje de cobertura del set, que llamaremos **Ø**), y **ii**) el porcentaje de genes del conjunto de entrada asociado a cada categoría de la definición proporcionada (representatividad de cada término en el set, que llamaremos **&**). Utilizamos este programa con dos finalidades:

a) En primer lugar, lo utilizamos para comprobar si los subconjuntos de proteínas que fueron asignados por el clasificador a una clase dada estaban en realidad bien descritos por la definición de la clase en cuestión. Es decir si los subconjuntos de la clase 1 estaban bien descritos por la definición del set1, los de la clase 2 por la definición del set 2 y así sucesivamente.

b) En segundo lugar para probar que la definición de una clase no permitiera hacer una descripción adecuada de los subconjuntos de proteínas asignados a otras clases distintas. Es decir, para comprobar que las definiciones de cada una de las clases fueran suficientemente excluyentes.

Los resultados obtenidos para las muestras aleatorias de 500 proteínas para el inciso (a) se presentan en las Tablas I.a-i, correspondientes a cada una de las 9 clases previamente definidas y a los tres subconjuntos asignados por nuestra red de Kohonen a cada una de estas clases. En el primer renglón de las tablas está el identificador del conjunto, en el segundo renglón el número de genes asociados a ese conjunto y entre paréntesis el tamaño de la muestra original a partir de la cual se obtuvo este conjunto, en el tercer renglón se presenta el porcentaje de cobertura (**Ø**), que nos indica el porcentaje de genes en el subconjunto cuya función se encuentra descrita por la definición de la clase, y por último, cada uno de los renglones restantes indican los porcentajes de genes de cada conjunto que pertenecen a los distintos términos de GO que describen la función de la clase (**&**).

De las Tablas I a-i puede verse que los valores de **Ø** son en su mayoría superiores al 80%, lo que indica que menos del 20% de los genes de cada conjunto presentan una función que no se encuentra descrita en la definición de la clase. Sin embargo, aún más importante que este parámetro, es el hecho de que la distribución porcentual de cada término (**&**) se ajuste a la distribución definida para la clase. Con la finalidad de comparar las distribuciones de la definición de la clase con las obtenidas para cada uno de los conjuntos (sets) asignados a dicha clase aplicamos una prueba estadística pareada no paramétrica, la prueba de Wilcoxon de rangos signados, para ver si las distribuciones (**&**) eran distintas. Los resultados de dicha prueba se presentan en las tablas II a y b. Como puede notarse, en todos los casos encontramos que las distribuciones no fueron distintas entre la definición de la clase y los conjuntos asignados a dicha clase.

Hasta aquí, podemos concluir que las definiciones de cada una de nuestras 9 clases nos permiten establecer un conjunto de funciones probables para cerca del 80% de los genes asignados por nuestro clasificador a dichas clases y adicionalmente, nuestra definición nos permite establecer de manera bastante acertada un porcentaje de probabilidad para cada una de las funciones asociadas a la clase.

Sin embargo, el análisis no puede estar completo si no hacemos la prueba inversa, es decir aplicar el criterio descrito en el inciso (**b**), para mostrar que la definición de una clase no permite hacer una descripción adecuada de conjuntos de genes asociados a otras clases distintas. Con esta finalidad, utilizamos el programa *ajuste_definicion.pl,* utilizando la definición de una clase y los conjuntos asignados a

todas las demás clases, para ver si las definiciones son lo suficientemente excluyentes como para poder discriminar entre conjuntos de genes pertenecientes a otras clases. Los resultados obtenidos para las muestras aleatorias de 500 proteínas se presentan en las tablas III.a-i.

Tabla I.a.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a la clase 1 obtenidos de muestras aleatorias de 500 genes, en relación a la definición de la clase 1 generada a partir de la clasificación original del genoma completo de S. *cerevisiae*.

Conjunto		**def_set1**	**set1_1**	**set1_2**	**set1_3**
Tamaño conjunto (muestra)		1116 (6700)	66 (500)	79 (500)	70 (500)
% Cobertura (Ø)		88%	89%	87%	87%
GO:0019538	Protein metabolism	21.58%	22.22%	30.00%	20.00%
GO:0006082	Organic acid metabolism	18.61%	18.52%	18.33%	20.00%
GO:0009308	Amine metabolism	13.47%	12.96%	10.00%	15.56%
GO:0006519	Aminoacid and derivative metabolism	12.79%	12.96%	10.00%	13.33%
GO:0006066	Alcohol metabolism	8.33%	11.11%	5.00%	4.44%
GO:0006950	Response to stress	7.42%	9.26%	6.67%	4.44%
GO:0006629	Lipid metabolism	6.51%	5.56%	10.00%	2.22%
GO:0006732	Coenzyme metabolism	5.71%	5.56%	8.33%	8.89%
GO:0006766	Vitamin metabolism	5.59%	9.26%	3.33%	2.22%
GO:0009117	Nucleotide metabolism	4.91%	9.26%	5.00%	6.67%
GO:0046483	Heterocycle metabolism	4.57%	3.70%	3.33%	8.89%
GO:0006092	Main pathways of carbohydrate metabolism	4.11%	3.70%	3.33%	4.44%
GO:0006811	Ion transport	4.00%	1.85%	10.00%	2.22%
GO:0006886	Intracellular protein transport	3.77%	0.00%	1.67%	4.44%
GO:0015849	Organic acid transport	3.54%	1.85%	5.00%	2.22%
GO:0006725	Aromatic compound metabolism	3.42%	3.70%	1.67%	2.22%
GO:0015837	Amine transport	3.42%	1.85%	3.33%	2.22%
GO:0048193	Golgi vesicle transport	3.20%	1.85%	5.00%	4.44%
GO:0006790	Sulfur metabolism	3.08%	1.85%	3.33%	4.44%
GO:0006399	tRNA metabolism	2.63%	0.00%	0.00%	4.44%
GO:0045333	Cellular respiration	2.40%	1.85%	3.33%	4.44%
GO:0008643	Carbohydrate transport	2.28%	3.70%	1.67%	2.22%
GO:0007005	Mitochondrion organization and biogenesis	2.05%	1.85%	3.33%	2.22%
GO:0006366	Transcription from RNA polymerase II promoter	1.94%	0.00%	3.33%	0.00%
GO:0030435	Sporulation	1.94%	1.85%	1.67%	4.44%
GO:0006897	Endocytosis	1.94%	5.56%	1.67%	2.22%
GO:0006364	RNA processing	1.71%	3.70%	0.00%	0.00%
GO:0007165	Signal transduction	1.71%	1.85%	0.00%	2.22%
GO:0006800	Oxygen and reactive oxygen species metabolism	1.48%	3.70%	0.00%	2.22%
GO:0006913	Nuclecytoplasmic transport	1.48%	0.00%	0.00%	0.00%
GO:0006113	Fermentation	1.37%	1.85%	0.00%	2.22%
GO:0006081	Aldehide metabolism	1.37%	1.85%	0.00%	0.00%
GO:0008219	Cell death	1.37%	0.00%	3.33%	0.00%
GO:0006312	Mitotic recombination	1.26%	1.85%	3.33%	0.00%
GO:0016311	Dephosphorylation	1.26%	0.00%	3.33%	2.22%
GO:0006073	Glucan metabolism	1.03%	0.00%	3.33%	2.22%
GO:0042493	Response to drug	0.91%	3.70%	1.67%	6.67%
GO:0006730	One-carbon compound metabolism	0.91%	1.85%	3.33%	0.00%
GO:0006118	Electron transport	0.91%	0.00%	0.00%	4.44%
GO:0006401	RNA catabolism	0.68%	1.85%	0.00%	0.00%
GO:0006352	Transcription initiation	0.68%	0.00%	0.00%	0.00%
GO:0006383	Transcription from RNA polymerase III promoter	0.68%	0.00%	1.67%	0.00%
GO:0030004	Monovalent inorganic cation homeostasis	0.57%	0.00%	0.00%	0.00%

Tabla I.b.- Porcentaje de cobertura y representatividad de los conjuntos asignados a la clase 2, según la definición de la misma.

Conjunto		**def_set2**	**set2_1**	**set2_2**	**set2_3**
Tamaño conjunto (muestra)		1446(6700)	119 (500)	93 (500)	107 (500)
% Cobertura (Ø)		89%	95%	84%	88%
GO:0006139	Nucleobase, nucleoside, nucleotide and nucleic acid metabolism	43.64%	40.43%	39.13%	46.99%
GO:0016043	Cell organization and biogénesis	36.25%	35.11%	36.23%	30.12%
GO:0050896	Response to stimulus	16.05%	19.15%	5.80%	20.48%
GO:0006464	Protein modification	15.42%	13.83%	10.14%	10.84%
GO:0007049	Cell cycle	12.89%	12.77%	15.94%	15.66%
GO:0006793	Phosphorus metabolism	6.31%	4.26%	5.80%	2.41%
GO:0007154	Cell comunication	5.95%	6.38%	2.90%	1.20%
GO:0007059	Chromosome segregation	3.43%	2.13%	4.35%	3.61%
GO:0000746	Conjugation	3.34%	4.26%	1.45%	2.41%
GO:0043037	Translation	2.98%	5.32%	4.35%	6.02%
GO:0030437	Sporulation (sensu Fungi)	2.80%	3.19%	0.00%	2.41%
GO:0009101	Glycoprotein biosíntesis	2.43%	2.13%	0.00%	1.20%
GO:0006944	Membrane fusion	1.53%	1.06%	2.90%	0.00%
GO:0006887	Exocytosis	1.35%	1.06%	1.45%	1.20%
GO:0008202	Steroid metabolism	1.26%	1.06%	1.45%	0.00%
GO:0009108	Coenzyme biosíntesis	1.17%	1.06%	1.45%	1.20%
GO:0006112	Energy reserve metabolism	0.81%	0.00%	1.45%	2.41%
GO:0006752	Group transfer coenzyme metabolism	0.81%	2.13%	0.00%	2.41%
GO:0046165	Alcohol biosíntesis	0.63%	1.06%	0.00%	0.00%
GO:0048278	Vesicle docking	0.54%	0.00%	1.45%	0.00%
GO:0006515	Misfolded or incompletely synthesized protein catabolism	0.45%	1.06%	0.00%	0.00%

Tabla I.c.- Pporcentaje de cobertura y representatividad de los conjuntos asignados a la clase 3, según la definición de la misma.

Conjunto		**def_set3**	**set3_1**	**set3_2**	**set3_3**
Tamaño conjunto (muestra)		613 (6700)	39 (500)	46 (500)	50 (500)
% Cobertura (Ø)		86%	81%	97%	79%
GO:0050896	Response to stimulus	16.27%	15.62%	18.18%	20.51%
GO:0019219	Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	16.05%	15.62%	18.18%	15.38%
GO:0007275	Development	15.84%	18.75%	15.15%	25.64%
GO:0007049	Cell cycle	15.18%	25.00%	9.09%	12.82%
GO:0007010	Cytoskeleton organization and biogénesis	12.58%	9.38%	18.18%	15.38%
GO:0007154	Cell communication	11.50%	6.25%	15.15%	15.38%
GO:0006313	DNA transposition	9.11%	21.88%	3.03%	7.69%
GO:0016192	Vesicle-mediated transport	8.03%	6.25%	0.00%	0.00%
GO:0007047	Cell wall organization and biogenesis	7.81%	9.38%	9.09%	15.38%
GO:0040007	Growth	7.38%	0.00%	9.09%	10.26%
GO:0006796	Phosphate metabolism	4.77%	6.25%	9.09%	0.00%
GO:0000746	Conjugation	4.56%	9.38%	3.03%	7.69%
GO:0006913	Nucleocytoplasmic transport	3.04%	0.00%	0.00%	5.13%
GO:0008219	Cell death	2.60%	0.00%	3.03%	5.13%
GO:0006073	Glucan metabolism	2.39%	6.25%	3.03%	2.56%
GO:0006997	Nuclear organization and biogénesis	2.39%	3.12%	0.00%	2.56%
GO:0006812	Cation transport	2.17%	0.00%	3.03%	0.00%
GO:0006402	mRNA catabolism	2.17%	0.00%	0.00%	0.00%
GO:0006644	Phospholipid metabolism	1.74%	3.12%	3.03%	5.13%
GO:0006875	Metal ion homeostasis	1.74%	3.12%	0.00%	0.00%
GO:0007028	Cytoplasm organization and biogénesis	1.74%	0.00%	0.00%	0.00%
GO:0006112	Energy reserve metabolism	1.30%	0.00%	3.03%	0.00%
GO:0048308	Organelle inheritance	1.30%	0.00%	3.03%	0.00%
GO:0031123	RNA 3'-end processing	1.08%	0.00%	0.00%	0.00%
GO:0051049	Regulation of transport	1.08%	3.12%	0.00%	2.56%
GO:0007109	Cytokinesis, completion of separation	0.65%	0.00%	0.00%	0.00%

Tabla I.d.- Porcentaje de cobertura y representatividad de los conjuntos asignados a la clase 4, según la definición de la misma.

Conjunto		**def_set4**	**set4_1**	**set4_2**	**set4_3**
Tamaño conjunto (muestra)		593 (6700)	41 (500)	42 (500)	38 (500)
% Cobertura (Ø)		83%	87%	88%	100%
GO:0006412	Protein biosíntesis	15.67%	13.04%	11.76%	40.00%
GO:0006396	RNA processing	11.67%	13.04%	23.53%	10.00%
GO:0006366	Transcription from RNA polymerase II promoter	10.00%	8.70%	5.88%	10.00%
GO:0042254	Ribosome biogenesis and assembly	9.67%	8.70%	17.65%	0.00%
GO:0006950	Response to stress	9.00%	8.70%	11.76%	15.00%
GO:0007126	Meiosis	4.33%	8.70%	5.88%	0.00%
GO:0009117	Nucleotide metabolism	4.00%	0.00%	0.00%	5.00%
GO:0045333	Cellular respiration	3.67%	0.00%	0.00%	0.00%
GO:0006811	Ion transport	3.67%	0.00%	0.00%	5.00%
GO:0006725	Aromatic compound metabolism	3.33%	0.00%	5.88%	0.00%
GO:0006457	Protein holding	3.00%	0.00%	0.00%	10.00%
GO:0045814	Negative regulation of gene expression, epigenetic	3.00%	0.00%	5.88%	10.00%
GO:0007047	Cell wall organization and biogenesis	3.00%	8.70%	0.00%	5.00%
GO:0006511	Ubiquitin-dependent protein catabolism	2.67%	4.35%	0.00%	0.00%
GO:0005996	Monosaccharide metabolism	2.33%	4.35%	0.00%	0.00%
GO:0030005	Di-, tri-valent inorganic cation homeostasis	2.33%	4.35%	5.88%	0.00%
GO:0007020	Microtubule nucleation	2.33%	0.00%	0.00%	5.00%
GO:0006875	Metal ion homeostasis	2.33%	4.35%	5.88%	0.00%
GO:0051028	mRNA transport	2.33%	0.00%	11.76%	0.00%
GO:0006626	Protein targeting to mitochondrion	2.00%	4.35%	5.88%	0.00%
GO:0006753	Nucleoside phosphate metabolism	2.00%	0.00%	0.00%	5.00%
GO:0007017	Microtubule-based process	1.67%	4.35%	5.88%	0.00%
GO:0043094	Metabolic compound salvage	1.67%	0.00%	0.00%	0.00%
GO:0006800	Oxygen and reactive oxygen species metabolism	1.67%	0.00%	0.00%	5.00%
GO:0006289	Nucleotide-excision repair	1.33%	0.00%	0.00%	0.00%
GO:0006100	Tricarboxylic acid cycle intermediate metabolism	1.33%	0.00%	0.00%	0.00%
GO:0006360	Transcription from RNA polymerase I promoter	1.33%	4.35%	0.00%	0.00%
GO:0006631	Fatty acid metabolism	1.33%	0.00%	0.00%	0.00%
GO:0006944	Membrane fusion	1.33%	0.00%	0.00%	0.00%
GO:0006914	Autophagy	1.00%	4.35%	0.00%	0.00%
GO:0008535	Cytochrome c oxidase complex assembly	1.00%	0.00%	0.00%	0.00%
GO:0030503	Regulation of cell redox homeostasis	1.00%	4.35%	0.00%	5.00%
GO:0007121	Bipolar bud site selection	1.00%	0.00%	5.88%	0.00%
GO:0007050	Cell cycle arrest	1.00%	4.35%	0.00%	0.00%
GO:0051052	Regulation of DNA metabolism	1.00%	0.00%	0.00%	0.00%
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	1.00%	0.00%	5.88%	0.00%
GO:0000002	Mitochondrial genome maintenance	0.67%	4.35%	0.00%	0.00%

Tabla I.e.- Porcentaje de cobertura y representatividad de los conjuntos asignados a la clase 5, según la definición de la misma.

Conjunto		**def_set5**	**set5_1**	**set5_2**	**set5_3**
Tamaño conjunto (muestra)		611 (6700)	40 (500)	37 (500)	44 (500)
% Cobertura (Ø)		91%	90%	89%	97%
GO:0019538	Protein metabolism	35.00%	38.71%	39.29%	33.33%
GO:0016070	RNA metabolism	19.09%	16.13%	25.00%	16.67%
GO:0007028	Cytoplasm organization and biogénesis	13.41%	0.00%	17.86%	10.00%
GO:0006351	Transcription, DNA-dependent	12.73%	12.90%	3.57%	10.00%
GO:0006325	Establishment and/or maintenance of chromatin architecture	7.50%	3.23%	0.00%	3.33%
GO:0006605	Protein targeting	7.27%	6.45%	0.00%	13.33%
GO:0016192	Vesicle-mediated transport	5.68%	3.23%	3.57%	10.00%
GO:0006091	Generation of precursor metabolites and energy	5.23%	6.45%	7.14%	3.33%
GO:0000087	M phase of mitotic cell cycle	4.09%	0.00%	3.57%	10.00%
GO:0007059	Chromosome segregation	4.09%	6.45%	0.00%	6.67%
GO:0006974	Response to DNA damage stimulus	3.86%	0.00%	3.57%	13.33%
GO:0006800	Oxygen and reactive oxygen species metabolism	2.95%	6.45%	3.57%	0.00%
GO:0007005	Mitochondrion organization and biogénesis	2.73%	6.45%	0.00%	0.00%
GO:0030468	Establishment of cell polarity (sensu Fungi)	2.73%	0.00%	3.57%	3.33%
GO:0006261	DNA-dependent DNA replication	2.50%	0.00%	7.14%	3.33%
GO:0030437	Sporulation (sensu Fungi)	2.27%	0.00%	3.57%	6.67%
GO:0050658	RNA transport	2.05%	6.45%	0.00%	0.00%
GO:0007114	Cell budding	2.05%	0.00%	0.00%	3.33%
GO:0006732	Coenzyme metabolism	1.59%	3.23%	3.57%	0.00%
GO:0007127	Meiosis	1.59%	0.00%	3.57%	3.33%
GO:0007264	Small GTPase mediated signal transduction	1.36%	0.00%	0.00%	3.33%
GO:0000086	G2/M transition of mitotic cell cycle	0.91%	0.00%	3.57%	0.00%
GO:0006109	Regulation of carbohydrate metabolism	0.91%	3.23%	3.57%	0.00%
GO:0007033	Vacuole organization and biogénesis	0.91%	0.00%	3.57%	0.00%
GO:0000067	DNA replication and chromosome cycle	0.23%	0.00%	0.00%	0.00%

Tabla I.f.- Porcentaje de cobertura y representatividad de los conjuntos asignados a la clase 6, según la definición de la misma.

Conjunto		**def_set6**	**set6_1**	**set6_2**	**set6_3**
Tamaño conjunto (muestra)		1018 (6700)	74 (500)	85 (500)	63 (500)
% Cobertura (Ø)		90%	94%	100%	96%
GO:0006810	Transport	31.79%	29.17%	34.62%	37.78%
GO:0006996	Organelle organization and biogénesis	19.69%	25.00%	19.23%	20.00%
GO:0006464	Protein modification	13.64%	12.50%	7.69%	24.44%
GO:0006629	Lipid metabolism	11.81%	18.75%	11.54%	11.11%
GO:0009628	Response to abiotic stimulus	6.05%	4.17%	5.77%	6.67%
GO:0000003	Reproduction	5.91%	4.17%	9.62%	4.44%
GO:0006355	Regulation of transcription, DNA-dependent	5.63%	8.33%	5.77%	2.22%
GO:0006082	Organic acid metabolism	4.78%	4.17%	3.85%	4.44%
GO:0030003	Cation homeostasis	4.22%	4.17%	3.85%	4.44%
GO:0007165	Signal transduction	3.80%	2.08%	5.77%	8.89%
GO:0009101	Glycoprotein biosíntesis	3.66%	4.17%	0.00%	15.56%
GO:0006511	Ubiquitin-dependent protein catabolism	3.52%	6.25%	0.00%	6.67%
GO:0000074	Regulation of progression through cell cycle	3.23%	4.17%	5.77%	2.22%
GO:0042158	Lipoprotein biosíntesis	3.23%	4.17%	0.00%	0.00%
GO:0008380	RNA splicing	2.95%	2.08%	3.85%	0.00%
GO:0006399	tRNA metabolism	2.81%	2.08%	1.92%	0.00%
GO:0006461	Protein complex assembly	2.39%	2.08%	5.77%	2.22%
GO:0045333	Cellular respiration	2.25%	2.08%	0.00%	2.22%
GO:0000910	Cytokinesis	2.25%	2.08%	1.92%	4.44%
GO:0006944	Membrane fusion	1.97%	0.00%	1.92%	2.22%
GO:0006281	DNA repair	1.97%	0.00%	1.92%	0.00%
GO:0016125	Sterol metabolism	1.69%	2.08%	1.92%	0.00%
GO:0016044	Membrane organization and biogénesis	1.55%	2.08%	1.92%	0.00%
GO:0006914	Autophagy	1.41%	0.00%	1.92%	0.00%
GO:0000070	Mitotic sister chromatid segregation	1.41%	2.08%	1.92%	0.00%
GO:0007131	Meiotic recombination	1.13%	0.00%	3.85%	0.00%
GO:0043414	Biopolymer methylation	0.98%	0.00%	0.00%	0.00%
GO:0000271	Polysaccharid biosíntesis	0.84%	0.00%	1.92%	2.22%
GO:0006270	DNA replication initiation	0.84%	0.00%	0.00%	0.00%
GO:0006515	Misfolded or incompletely synthesized protein catabolism	0.84%	4.17%	0.00%	2.22%
GO:0006743	Ubiquinone metabolism	0.70%	0.00%	0.00%	0.00%
GO:0007062	Syster chromatide cohesión	0.70%	2.08%	1.92%	2.22%
GO:0006267	Pre-replicative complex formation and maintenance	0.70%	0.00%	1.92%	0.00%
GO:0006308	DNA catabolism	0.56%	0.00%	0.00%	0.00%
GO:0007091	Mitotic metaphase/anaphase transition	0.56%	0.00%	0.00%	0.00%
GO:0000080	G1 phase of mitotic cell cycle	0.56%	2.08%	0.00%	2.22%
GO:0042168	Heme metabolism	0.56%	0.00%	3.85%	0.00%
GO:0006298	Mismatch repair	0.56%	2.08%	1.92%	2.22%
GO:0000288	mRNA catabolism, deadenylylation-dependent decay	0.56%	2.08%	0.00%	0.00%
GO:0042401	Biogenic amine biosíntesis	0.56%	0.00%	0.00%	0.00%

Tabla I.g.- Porcentaje de cobertura y representatividad de los conjuntos asignados a la clase 7, según la definición de la misma.

Conjunto		**def_set7**	**set7_1**	**set7_2**	**set7_3**
Tamaño conjunto (muestra)		222 (6700)	19 (500)	13 (500)	14 (500)
% Cobertura (Ø)		70%	100%	100%	100%
GO:0006412	Protein biosíntesis	30.43%	0.00%	0.00%	50.00%
GO:0010038	Response to metal ion	13.04%	50.00%	0.00%	50.00%
GO:0015986	ATP synthesis coupled proton transport	8.70%	0.00%	0.00%	0.00%
GO:0006812	Cation transport	8.70%	50.00%	100.00%	0.00%
GO:0006091	Oxidative phosphorilation	4.35%	0.00%	0.00%	0.00%
GO:0006605	Protein targeting	4.35%	0.00%	0.00%	0.00%

Tabla I.h.- Porcentaje de cobertura y representatividad de los conjuntos asignados a la clase 8, según la definición de la misma.

Conjunto		**def_set8**	**set8_1**	**set8_2**	**set8_3**
Tamaño conjunto (muestra)		545 (6700)	45 (500)	41 (500)	54 (500)
% Cobertura (Ø)		82%	92%	88%	82%
GO:0019538	Protein metabolism	37.20%	28.00%	24.00%	32.35%
GO:0042254	Ribosome biogenesis and assembly	9.50%	8.00%	12.00%	8.82%
GO:0007010	Cytoskeleton organization and biogénesis	7.65%	0.00%	12.00%	0.00%
GO:0009628	Response to abiotic stimulus	6.33%	0.00%	12.00%	8.82%
GO:0006357	Regulation of transcription from RNA polymerase II promoter	5.54%	4.00%	8.00%	11.76%
GO:0045045	Secretory pathway	5.28%	16.00%	4.00%	8.82%
GO:0006323	DNA packaging	5.28%	12.00%	4.00%	5.88%
GO:0008380	RNA splicing	4.22%	8.00%	4.00%	8.82%
GO:0006119	Oxidative phosphorilation	3.96%	0.00%	8.00%	5.88%
GO:0048193	Golgi vesicle transport	3.69%	12.00%	4.00%	8.82%
GO:0006913	Nucleocytoplasmic transport	3.43%	12.00%	4.00%	0.00%
GO:0017038	Protein import	3.43%	4.00%	0.00%	0.00%
GO:0007059	Chromosome segregation	3.17%	4.00%	8.00%	5.88%
GO:0007032	Endosome organization and biogénesis	2.37%	4.00%	4.00%	2.94%
GO:0030471	Spindle pole body and microtubule cycle (sensu Fungi)	1.85%	0.00%	4.00%	2.94%
GO:0006818	Hydrogen transport	1.85%	0.00%	8.00%	2.94%
GO:0006092	Main pathways of carbohydrate metabolism	1.85%	4.00%	4.00%	0.00%
GO:0006383	Transcription from RNA polymerase III promoter	1.58%	0.00%	0.00%	0.00%
GO:0007034	Vacuolar transport	1.58%	4.00%	0.00%	2.94%
GO:0007088	Regulation of mitosis	1.32%	4.00%	4.00%	2.94%
GO:0007033	Vacuole organization and biogénesis	1.32%	0.00%	0.00%	2.94%
GO:0000096	Sulfur amino acid metabolism	1.06%	0.00%	0.00%	0.00%
GO:0030641	Hydrogen ion homeostasis	1.06%	0.00%	0.00%	0.00%

Tabla I.i.- Porcentaje de cobertura y representatividad de los conjuntos asignados a la clase 9, según la definición de la misma.

Conjunto		**def_set9**	**set9_1**	**set9_2**	**set9_3**
Tamaño conjunto (muestra)		536 (6700)	39 (500)	41 (500)	40 (500)
% Cobertura (Ø)		61%	100%	80%	100%
GO:0006412	Protein biosíntesis	28.95%	100%	40.00%	100.00%
GO:0006626	Protein targeting to mitochondrion	7.89%	0.00%	20.00%	0.00%
GO:0000749	Response to pheromone during conjugation with cellular fusion	7.89%	0.00%	20.00%	0.00%
GO:0016197	Endosome transport	7.89%	0.00%	0.00%	0.00%
GO:0045333	Cellular respiration	5.26%	0.00%	20.00%	50.00%
GO:0046467	Membrane lipid biosíntesis	5.26%	0.00%	0.00%	50.00%
GO:0000028	Ribosomal small subunit assembly and maintenance	5.26%	0.00%	0.00%	0.00%
GO:0006450	Regulation of translational fidelity	5.26%	0.00%	0.00%	0.00%
GO:0007105	Cytokinesis, site selection	2.63%	0.00%	0.00%	0.00%

Tabla II.a.- Resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de las clases 1-5 y los subconjuntos asignados a esas mismas clases a partir de muestras aleatorias de 500 genes.

Wilcoxon signed rank test			
	Def1_Vs_set1_1	**Def1_Vs_set1_2**	**Def1_Vs_set1_3**
P value	0.333	0.427	0.4988
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	509.0 , -437.0	488.5 , -457.5	473.0 , -473.0
Sum of signed ranks (W)	72	31	0
How effective was the pairing?			
rs (Spearman, Approximation)	0.7282	0.7477	0.7046
P Value	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***
Was the pairing significantly effective?	Yes	Yes	Yes
	Def2_Vs_set2_1	**Def2_Vs_set2_2**	**Def2_Vs_set2_3**
P value	0.2981	0.1104	0.2096
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	131.0 , -100.0	151.0 , -80.00	139.0 , -92.00
Sum of signed ranks (W)	31	71	47
How effective was the pairing?			
rs (Spearman, Approximation)	0.9208	0.8399	0.8208
P Value	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***
Was the pairing significantly effective?	Yes	Yes	Yes
	Def3_Vs_set3_1	**Def3_Vs_set3_2**	**Def3_Vs_set3_3**
P value	0.0663	0.427	0.2137
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	235.0 , -116.0	183.0 , -168.0	144.0 , -207.0
Sum of signed ranks (W)	119	15	-63
How effective was the pairing?			
rs (Spearman, Approximation)	0.8137	0.7792	0.7955
P Value	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***
Was the pairing significantly effective?	Yes	Yes	Yes
	Def4_Vs_set4_1	**Def4_Vs_set4_2**	**Def4_Vs_set4_3**
P value	0.4985	0.4804	0.184
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	352.0 , -351.0	348.0 , -355.0	390.5 , -275.5
Sum of signed ranks (W)	1	-7	115
How effective was the pairing?			
rs (Spearman, Approximation)	0.3874	0.4932	0.5465
P Value	0.0089	0.001	0.0002
P value summary	**	***	***
Was the pairing significantly effective?	Yes	Yes	Yes
	Def5_Vs_set5_1	**Def5_Vs_set5_2**	**Def5_Vs_set5_3**
P value	0.1395	0.4332	0.4866
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	203.0 , -122.0	156.0 , -169.0	161.0 , -164.0
Sum of signed ranks (W)	81	-13	-3
How effective was the pairing?			
rs (Spearman, Approximation)	0.5711	0.3715	0.7878
P Value (one tailed)	0.0014	0.0337	P<0.0001
P value summary	**	*	***
Was the pairing significantly effective?	Yes	Yes	Yes

Tabla II.b.- Resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de las clases 6-9 y los subconjuntos asignados a esas mismas clases a partir de muestras aleatorias de 500 genes.

Wilcoxon signed rank test			
	Def6_Vs_set6_1	**Def6_Vs_set6_2**	**Def6_Vs_set6_3**
P value	0.3013	0.4585	0.2693
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	449.0 , -371.0	418.0 , -402.0	456.0 , -364.0
Sum of signed ranks (W)	78	16	92
How effective was the pairing?			
rs (Spearman, Approximation)	0.7813	0.6769	0.704
P Value	P<0.0001	P<0.0001	P<0.0001
P value summary	***	***	***
Was the pairing significantly effective?	Yes	Yes	Yes
	Def7_Vs_set7_1	**Def7_Vs_set7_2**	**Def7_Vs_set7_3**
P value	0.5	0.2188	0.5
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	10.00 , -11.00	15.00 , -6.000	10.00 , -11.00
Sum of signed ranks (W)	-1	9	-1
How effective was the pairing?			
rs (Spearman, Approximation)	0.3198	0	0.8528
P Value	0.2819	0.5	0.0167
P value summary	Ns	ns	*
Was the pairing significantly effective?	No	No	Yes
	Def8_Vs_set8_1	**Def8_Vs_set8_2**	**Def8_Vs_set8_3**
P value	0.3834	0.0733	0.191
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	128.0 , -148.0	90.00 , -186.0	109.0 , -167.0
Sum of signed ranks (W)	-20	-96	-58
How effective was the pairing?			
rs (Spearman, Approximation)	0.4883	0.7753	0.6549
P Value	0.009	P<0.0001	0.0003
P value summary	**	***	***
Was the pairing significantly effective?	Yes	Yes	Yes
	Def9_Vs_set9_1	**Def9_Vs_set9_2**	**Def9_Vs_set9_3**
P value	0.0645	0.2129	0.4551
Are medians signif. different? (P < 0.05)	No	No	No
Sum of positive, negative ranks	36.00 , -9.000	15.00 , -30.00	21.00 , -24.00
Sum of signed ranks (W)	27	-15	-3
How effective was the pairing?			
rs (Spearman, Approximation)	0.5828	0.6741	0.212
P Value	0.0498	0.0232	0.292
P value summary	*	*	ns
Was the pairing significantly effective?	Yes	Yes	No

Tabla III.a.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 2-9 obtenidos de muestras aleatorias de 500 genes, en relación a la definición de la clase 1 generada a partir de la clasificación original del genoma completo de S. *cerevisiae*.

Conjunto	**def_set1**	**set2**	**set3**	**set4**	**set5**	**set6**	**set7**	**set8**	**set9**
Tamaño (Muestra)	1116 (6700)	119 (500)	39 (500)	41 (500)	40 (500)	74 (500)	19 (500)	45 (500)	39 (500)
Cobertura	**88%**	**71%**	**61%**	**65%**	**71%**	**79%**	**50%**	**80%**	**0%**
GO:0019538	21.58%	27.66%	10.71%	17.39%	38.71%	20.83%	0.00%	28.00%	0.00%
GO:0006082	18.61%	8.51%	0.00%	0.00%	9.68%	4.17%	0.00%	4.00%	0.00%
GO:0009308	13.47%	6.38%	0.00%	0.00%	6.45%	0.00%	0.00%	0.00%	0.00%
GO:0006519	12.79%	5.32%	0.00%	0.00%	6.45%	0.00%	0.00%	0.00%	0.00%
GO:0006066	8.33%	2.13%	3.57%	4.35%	3.23%	6.25%	0.00%	4.00%	0.00%
GO:0006950	7.42%	13.83%	7.14%	8.70%	12.90%	6.25%	0.00%	8.00%	0.00%
GO:0006629	6.51%	4.26%	0.00%	0.00%	0.00%	18.75%	0.00%	0.00%	0.00%
GO:0006732	5.71%	2.13%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%	0.00%
GO:0006766	5.59%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0009117	4.91%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0046483	4.57%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006092	4.11%	1.06%	0.00%	0.00%	6.45%	2.08%	0.00%	4.00%	0.00%
GO:0006811	4.00%	1.06%	7.14%	0.00%	0.00%	2.08%	50%	4.00%	0.00%
GO:0006886	3.77%	6.38%	7.14%	8.70%	6.45%	6.25%	0.00%	8.00%	0.00%
GO:0015849	3.54%	0.00%	0.00%	4.35%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0006725	3.42%	2.13%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%	0.00%
GO:0015837	3.42%	0.00%	0.00%	4.35%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0048193	3.20%	1.06%	3.57%	0.00%	0.00%	4.17%	0.00%	12.00%	0.00%
GO:0006790	3.08%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006399	2.63%	6.38%	0.00%	0.00%	3.23%	2.08%	0.00%	0.00%	0.00%
GO:0045333	2.40%	1.06%	0.00%	0.00%	3.23%	2.08%	0.00%	0.00%	0.00%
GO:0008643	2.28%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007005	2.05%	3.19%	3.57%	8.70%	6.45%	6.25%	0.00%	0.00%	0.00%
GO:0006366	1.94%	6.38%	7.14%	8.70%	9.68%	8.33%	0.00%	4.00%	0.00%
GO:0030435	1.94%	3.19%	3.57%	4.35%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0006897	1.94%	1.06%	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006364	1.71%	0.00%	0.00%	8.70%	0.00%	2.08%	0.00%	4.00%	0.00%
GO:0007165	1.71%	6.38%	10.71%	0.00%	0.00%	2.08%	0.00%	4.00%	0.00%
GO:0006800	1.48%	0.00%	0.00%	0.00%	6.45%	0.00%	0.00%	0.00%	0.00%
GO:0006913	1.48%	3.19%	0.00%	0.00%	6.45%	4.17%	0.00%	12.00%	0.00%
GO:0006113	1.37%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006081	1.37%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0008219	1.37%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0006312	1.26%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0016311	1.26%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006073	1.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0042493	0.91%	1.06%	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006730	0.91%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006118	0.91%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006401	0.68%	1.06%	0.00%	0.00%	0.00%	2.08%	0.00%	4.00%	0.00%
GO:0006352	0.68%	3.19%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%	0.00%
GO:0006383	0.68%	3.19%	3.57%	4.35%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0030004	0.57%	0.00%	0.00%	4.35%	0.00%	2.08%	0.00%	0.00%	0.00%

Tabla III.b.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1 y 3-9 en relación a la definición de la clase 2.

Conjunto	def_set2	set1	set3	set4	set5	set6	set7	set8	set9
Tamaño (muestra)	1446 (6700)	66 (500)	39 (500)	41 (500)	40 (500)	74 (500)	19 (500)	45 (500)	39 (500)
Cobertura	**89%**	**52%**	**88%**	**70%**	**68%**	**76%**	**50%**	**75%**	**0%**
GO:0006139	43.64%	14.81%	32.14%	30.43%	35.48%	20.83%	0.00%	28.00%	0.00%
GO:0016043	36.25%	18.52%	57.14%	34.78%	19.35%	39.58%	0.00%	48.00%	0.00%
GO:0050896	16.05%	12.96%	14.29%	13.04%	16.13%	8.33%	50.00%	8.00%	0.00%
GO:0006464	15.42%	3.70%	7.14%	0.00%	9.68%	12.50%	0.00%	4.00%	0.00%
GO:0007049	12.89%	0.00%	21.43%	13.04%	3.23%	10.42%	0.00%	4.00%	0.00%
GO:0006793	6.31%	1.85%	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007154	5.95%	1.85%	10.71%	0.00%	0.00%	2.08%	0.00%	4.00%	0.00%
GO:0007059	3.43%	0.00%	0.00%	4.35%	6.45%	4.17%	0.00%	4.00%	0.00%
GO:0000746	3.34%	0.00%	3.57%	4.35%	0.00%	4.17%	0.00%	0.00%	0.00%
GO:0043037	2.98%	1.85%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%	0.00%
GO:0030437	2.80%	1.85%	3.57%	4.35%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0009101	2.43%	0.00%	0.00%	0.00%	3.23%	4.17%	0.00%	0.00%	0.00%
GO:0006944	1.53%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0006887	1.35%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0008202	1.26%	1.85%	0.00%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0009108	1.17%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006112	0.81%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006752	0.81%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0046165	0.63%	1.85%	0.00%	0.00%	3.23%	2.08%	0.00%	4.00%	0.00%
GO:0048278	0.54%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006515	0.45%	1.85%	0.00%	0.00%	0.00%	4.17%	0.00%	0.00%	0.00%

Tabla III.c.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1-2 y 4-9 en relación a la definición de la clase 3.

Conjunto	def_set3	set1	set2	set4	set5	set6	set7	set8	set9
Tamaño (muestra)	613 (6700)	66 (500)	119 (500)	41 (500)	40 (500)	74 (500)	19 (500)	45 (500)	39 (500)
Cobertura	**86%**	**34%**	**66%**	**51%**	**39%**	**57%**	**100%**	**55%**	**0%**
GO:0050896	16.27%	12.96%	19.15%	13.04%	16.13%	8.33%	50.00%	8.00%	0.00%
GO:0019219	16.05%	0.00%	8.51%	8.70%	6.45%	8.33%	0.00%	4.00%	0.00%
GO:0007275	15.84%	3.70%	10.64%	4.35%	3.23%	2.08%	0.00%	8.00%	0.00%
GO:0007049	15.18%	0.00%	12.77%	13.04%	3.23%	10.42%	0.00%	4.00%	0.00%
GO:0007010	12.58%	1.85%	10.64%	4.35%	0.00%	12.50%	0.00%	0.00%	0.00%
GO:0007154	11.50%	1.85%	6.38%	0.00%	0.00%	2.08%	0.00%	4.00%	0.00%
GO:0006313	9.11%	0.00%	4.26%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0016192	8.03%	7.41%	5.32%	0.00%	3.23%	6.25%	0.00%	16.00%	0.00%
GO:0007047	7.81%	3.70%	7.45%	8.70%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0040007	7.38%	1.85%	6.38%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006796	4.77%	1.85%	4.26%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0000746	4.56%	0.00%	4.26%	4.35%	0.00%	4.17%	0.00%	0.00%	0.00%
GO:0006913	3.04%	0.00%	3.19%	0.00%	6.45%	4.17%	0.00%	12.00%	0.00%
GO:0008219	2.60%	0.00%	1.06%	0.00%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0006073	2.39%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006997	2.39%	0.00%	1.06%	0.00%	0.00%	2.08%	0.00%	4.00%	0.00%
GO:0006812	2.17%	1.85%	1.06%	0.00%	0.00%	2.08%	50.00%	0.00%	0.00%
GO:0006402	2.17%	1.85%	0.00%	0.00%	0.00%	2.08%	0.00%	4.00%	0.00%
GO:0006644	1.74%	1.85%	1.06%	0.00%	0.00%	12.50%	0.00%	0.00%	0.00%
GO:0006875	1.74%	0.00%	0.00%	4.35%	3.23%	2.08%	0.00%	0.00%	0.00%
GO:0007028	1.74%	5.56%	0.00%	8.70%	0.00%	2.08%	0.00%	8.00%	0.00%
GO:0006112	1.30%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0048308	1.30%	0.00%	1.06%	0.00%	3.23%	0.00%	0.00%	0.00%	0.00%
GO:0031123	1.08%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0051049	1.08%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007109	0.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Tabla III.d.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1-3 y 5-9 en relación a la definición de la clase 4.

Conjunto	**def_set4**	**set1**	**set2**	**set3**	**set5**	**set6**	**set7**	**set8**	**set9**
Tamaño (muestra)	593 (6700)	66 (500)	119 (500)	39 (500)	40 (500)	74 (500)	19 (500)	45 (500)	39 (500)
Cobertura	**83%**	**50%**	**53%**	**45%**	**57%**	**55%**	**50%**	**71%**	**0%**
GO:0006412	15.67%	9.26%	8.51%	0.00%	18.18%	10.42%	0.00%	16.00%	0.00%
GO:0006396	11.67%	3.70%	6.38%	3.57%	6.06%	4.17%	0.00%	8.00%	0.00%
GO:0006366	10.00%	0.00%	6.38%	7.14%	6.06%	8.33%	0.00%	4.00%	0.00%
GO:0042254	9.67%	5.56%	0.00%	3.57%	6.06%	2.08%	0.00%	8.00%	0.00%
GO:0006950	9.00%	9.26%	12.77%	7.14%	9.09%	6.25%	0.00%	8.00%	0.00%
GO:0007126	4.33%	0.00%	4.26%	10.71%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0009117	4.00%	9.26%	0.00%	0.00%	3.03%	0.00%	0.00%	0.00%	0.00%
GO:0045333	3.67%	1.85%	1.06%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0006811	3.67%	1.85%	1.06%	7.14%	0.00%	2.08%	50.00%	4.00%	0.00%
GO:0006725	3.33%	3.70%	2.13%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006457	3.00%	0.00%	0.00%	0.00%	3.03%	2.08%	0.00%	4.00%	0.00%
GO:0045814	3.00%	0.00%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007047	3.00%	3.70%	7.45%	14.29%	3.03%	0.00%	0.00%	0.00%	0.00%
GO:0006511	2.67%	1.85%	0.00%	3.57%	6.06%	6.25%	0.00%	4.00%	0.00%
GO:0005996	2.33%	7.41%	1.06%	3.57%	0.00%	4.17%	0.00%	4.00%	0.00%
GO:0030005	2.33%	0.00%	0.00%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0007020	2.33%	0.00%	0.00%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0006875	2.33%	0.00%	0.00%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0051028	2.33%	0.00%	2.13%	0.00%	0.00%	2.08%	0.00%	8.00%	0.00%
GO:0006626	2.00%	0.00%	1.06%	0.00%	3.03%	4.17%	0.00%	0.00%	0.00%
GO:0006753	2.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007017	1.67%	0.00%	3.19%	3.57%	0.00%	10.42%	0.00%	0.00%	0.00%
GO:0043094	1.67%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006800	1.67%	3.70%	0.00%	0.00%	3.03%	0.00%	0.00%	0.00%	0.00%
GO:0006289	1.33%	0.00%	3.19%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006100	1.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006360	1.33%	0.00%	0.00%	0.00%	3.03%	0.00%	0.00%	0.00%	0.00%
GO:0006631	1.33%	1.85%	1.06%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0006944	1.33%	1.85%	1.06%	0.00%	3.03%	0.00%	0.00%	4.00%	0.00%
GO:0006914	1.00%	0.00%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0008535	1.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0030503	1.00%	0.00%	0.00%	0.00%	3.03%	0.00%	0.00%	0.00%	0.00%
GO:0007121	1.00%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007050	1.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0051052	1.00%	0.00%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006614	1.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0000002	0.67%	1.85%	2.13%	0.00%	3.03%	2.08%	0.00%	0.00%	0.00%

Tabla III.e.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1-4 y 6-9 en relación a la definición de la clase 5.

Conjunto	**def_set5**	**set1**	**set2**	**set3**	**set4**	**set6**	**set7**	**set8**	**set9**
Tamaño (muestra)	611 (6700)	66 (500)	119 (500)	39 (500)	41 (500)	74 (500)	19 (500)	45 (500)	39 (500)
Cobertura	**91%**	**49%**	**75%**	**53%**	**64%**	**62%**	**0%**	**87%**	**0%**
GO:0019538	35.00%	22.22%	27.66%	10.71%	17.39%	20.83%	0.00%	28.00%	0.00%
GO:0016070	19.09%	3.70%	11.70%	3.57%	13.04%	6.25%	0.00%	12.00%	0.00%
GO:0007028	13.41%	5.56%	0.00%	3.57%	8.70%	2.08%	0.00%	8.00%	0.00%
GO:0006351	12.73%	0.00%	11.70%	14.29%	13.04%	8.33%	0.00%	4.00%	0.00%
GO:0006325	7.50%	0.00%	4.26%	3.57%	0.00%	2.08%	0.00%	12.00%	0.00%
GO:0006605	7.27%	0.00%	6.38%	3.57%	8.70%	4.17%	0.00%	8.00%	0.00%
GO:0016192	5.68%	7.41%	5.32%	10.71%	0.00%	6.25%	0.00%	16.00%	0.00%
GO:0006091	5.23%	7.41%	2.13%	0.00%	0.00%	4.17%	0.00%	4.00%	0.00%
GO:0000087	4.09%	0.00%	3.19%	7.14%	4.35%	6.25%	0.00%	4.00%	0.00%
GO:0007059	4.09%	0.00%	2.13%	0.00%	4.35%	4.17%	0.00%	4.00%	0.00%
GO:0006974	3.86%	0.00%	8.51%	0.00%	4.35%	4.17%	0.00%	4.00%	0.00%
GO:0006800	2.95%	3.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007005	2.73%	1.85%	3.19%	3.57%	8.70%	6.25%	0.00%	0.00%	0.00%
GO:0030468	2.73%	1.85%	3.19%	3.57%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0006261	2.50%	0.00%	3.19%	0.00%	4.35%	4.17%	0.00%	0.00%	0.00%
GO:0030437	2.27%	1.85%	3.19%	3.57%	4.35%	0.00%	0.00%	4.00%	0.00%
GO:0050658	2.05%	0.00%	2.13%	0.00%	0.00%	2.08%	0.00%	8.00%	0.00%
GO:0007114	2.05%	1.85%	1.06%	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006732	1.59%	5.56%	2.13%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007127	1.59%	0.00%	1.06%	3.57%	4.35%	0.00%	0.00%	0.00%	0.00%
GO:0007264	1.36%	1.85%	1.06%	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0000086	0.91%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006109	0.91%	0.00%	1.06%	0.00%	0.00%	2.08%	0.00%	0.00%	0.00%
GO:0007033	0.91%	0.00%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0000067	0.23%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Tabla III.f.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1-5 y 7-9 en relación a la definición de la clase 6.

Conjunto	**def_set6**	**set1**	**set2**	**set3**	**set4**	**set5**	**set7**	**set8**	**set9**
Tamaño (muestra)	1018 (6700)	66 (500)	119 (500)	39 (500)	41 (500)	40 (500)	19 (500)	45 (500)	39 (500)
Cobertura	**90%**	**65%**	**79%**	**79%**	**45%**	**68%**	**100%**	**75%**	**0%**
GO:0006810	31.79%	22.22%	7.23%	25.00%	5.00%	12.90%	50.00%	36.00%	0.00%
GO:0006996	19.69%	12.96%	22.89%	32.14%	25.00%	12.90%	0.00%	24.00%	0.00%
GO:0006464	13.64%	3.70%	10.84%	7.14%	5.00%	9.68%	0.00%	4.00%	0.00%
GO:0006629	11.81%	5.56%	3.61%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0009628	6.05%	11.11%	7.23%	10.71%	5.00%	9.68%	50.00%	0.00%	0.00%
GO:0000003	5.91%	3.70%	4.82%	10.71%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0006355	5.63%	0.00%	6.02%	10.71%	15.00%	3.23%	0.00%	4.00%	0.00%
GO:0006082	4.78%	18.52%	7.23%	0.00%	0.00%	9.68%	0.00%	4.00%	0.00%
GO:0030003	4.22%	0.00%	0.00%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%
GO:0007165	3.80%	1.85%	0.00%	10.71%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0009101	3.66%	0.00%	1.20%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%
GO:0006511	3.52%	1.85%	2.41%	3.57%	0.00%	3.23%	0.00%	4.00%	0.00%
GO:0000074	3.23%	0.00%	3.61%	7.14%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0042158	3.23%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0008380	2.95%	0.00%	2.41%	3.57%	10.00%	12.90%	0.00%	8.00%	0.00%
GO:0006399	2.81%	0.00%	6.02%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%
GO:0006461	2.39%	1.85%	3.61%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%
GO:0045333	2.25%	1.85%	1.20%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%
GO:0000910	2.25%	1.85%	0.00%	7.14%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006944	1.97%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0006281	1.97%	0.00%	9.64%	0.00%	0.00%	0.00%	0.00%	4.00%	0.00%
GO:0016125	1.69%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0016044	1.55%	0.00%	0.00%	0.00%	0.00%	3.23%	0.00%	0.00%	0.00%
GO:0006914	1.41%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0000070	1.41%	0.00%	2.41%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007131	1.13%	0.00%	3.61%	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0043414	0.98%	0.00%	1.20%	0.00%	5.00%	6.45%	0.00%	0.00%	0.00%
GO:0000271	0.84%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006270	0.84%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006515	0.84%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006743	0.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007062	0.70%	0.00%	2.41%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006267	0.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006308	0.56%	0.00%	1.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007091	0.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0000080	0.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0042168	0.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006298	0.56%	0.00%	1.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0000288	0.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0042401	0.56%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Tabla III.g.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1-6 y 8-9 en relación a la definición de la clase 7.

Conjunto	def_set7	set1	set2	set3	set4	set5	set6	set8	set9
Tamaño (muestra)	222 (6700)	66 (500)	119 (500)	39 (500)	41 (500)	40 (500)	74 (500)	45 (500)	39 (500)
Cobertura	**70%**	**18%**	**17%**	**6%**	**22%**	**27%**	**21%**	**27%**	**0%**
GO:0006412	30.43%	9.26%	8.51%	0.00%	13.04%	18.18%	10.42%	16.00%	0.00%
GO:0010038	13.04%	0.00%	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0015986	8.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006812	8.70%	1.85%	1.06%	3.57%	0.00%	0.00%	2.08%	0.00%	0.00%
GO:0006091	4.35%	7.41%	2.13%	0.00%	0.00%	0.00%	4.17%	4.00%	0.00%
GO:0006605	4.35%	0.00%	6.38%	3.57%	8.70%	9.09%	4.17%	8.00%	0.00%

Tabla III.h.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1-7 y 9 en relación a la definición de la clase 8.

Conjunto	def_set8	set1	set2	set3	set4	set5	set6	set7	set9
Tamaño (muestra)	545 (6700)	66 (500)	119 (500)	39 (500)	41 (500)	40 (500)	74 (500)	19 (500)	39 (500)
Cobertura	**82%**	**43%**	**65%**	**45%**	**64%**	**76%**	**57%**	**50%**	**0%**
GO:0019538	37.20%	22.22%	27.66%	10.71%	17.39%	33.33%	20.83%	0.00%	0.00%
GO:0042254	9.50%	5.56%	0.00%	3.57%	8.70%	6.06%	2.08%	0.00%	0.00%
GO:0007010	7.65%	1.85%	10.64%	14.29%	4.35%	3.03%	12.50%	0.00%	0.00%
GO:0009628	6.33%	11.11%	8.51%	10.71%	4.35%	3.03%	4.17%	50.00%	0.00%
GO:0006357	5.54%	0.00%	4.26%	7.14%	8.70%	6.06%	8.33%	0.00%	0.00%
GO:0045045	5.28%	1.85%	3.19%	3.57%	0.00%	12.12%	4.17%	0.00%	0.00%
GO:0006323	5.28%	0.00%	4.26%	3.57%	0.00%	6.06%	2.08%	0.00%	0.00%
GO:0008380	4.22%	0.00%	6.38%	3.57%	4.35%	6.06%	2.08%	0.00%	0.00%
GO:0006119	3.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0048193	3.69%	1.85%	1.06%	3.57%	0.00%	12.12%	4.17%	0.00%	0.00%
GO:0006913	3.43%	0.00%	3.19%	0.00%	0.00%	6.06%	4.17%	0.00%	0.00%
GO:0017038	3.43%	0.00%	2.13%	0.00%	4.35%	3.03%	2.08%	0.00%	0.00%
GO:0007059	3.17%	0.00%	2.13%	0.00%	4.35%	0.00%	4.17%	0.00%	0.00%
GO:0007032	2.37%	1.85%	3.19%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0030471	1.85%	0.00%	1.06%	3.57%	0.00%	0.00%	2.08%	0.00%	0.00%
GO:0006818	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006092	1.85%	3.70%	1.06%	0.00%	0.00%	0.00%	2.08%	0.00%	0.00%
GO:0006383	1.58%	0.00%	3.19%	3.57%	4.35%	0.00%	0.00%	0.00%	0.00%
GO:0007034	1.58%	0.00%	1.06%	0.00%	4.35%	3.03%	4.17%	0.00%	0.00%
GO:0007088	1.32%	0.00%	0.00%	0.00%	0.00%	3.03%	0.00%	0.00%	0.00%
GO:0007033	1.32%	0.00%	1.06%	0.00%	0.00%	3.03%	0.00%	0.00%	0.00%
GO:0000096	1.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0030641	1.06%	0.00%	0.00%	0.00%	4.35%	0.00%	2.08%	0.00%	0.00%

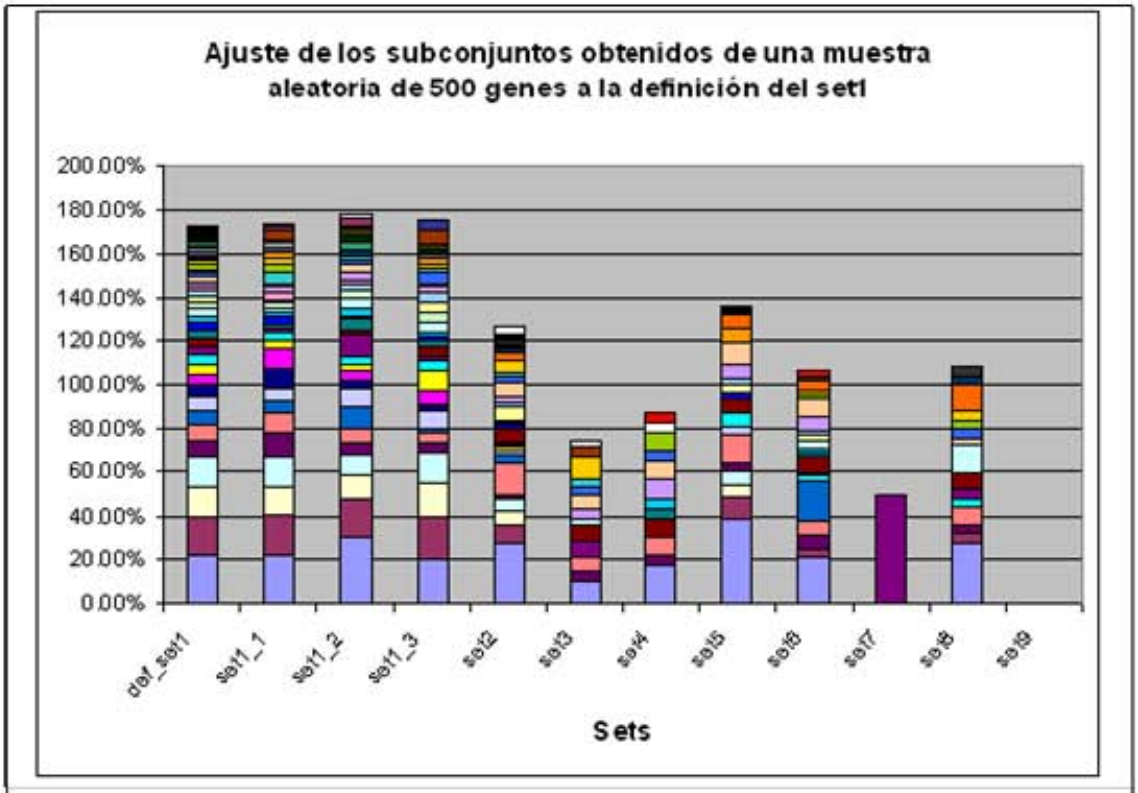
Tabla III.i.- Porcentaje de cobertura y representatividad de cada término de GO en los conjuntos asignados a las clases 1-8 en relación a la definición de la clase 9.

Conjunto	def_set9	set1	set2	set3	set4	set5	set6	set7	set8
Tamaño (muestra)	536 (6700)	66 (500)	119 (500)	39 (500)	41 (500)	40 (500)	74 (500)	19 (500)	45 (500)
Cobertura	**61%**	**17%**	**16%**	**7%**	**22%**	**28%**	**26%**	**0%**	**20%**
GO:0006412	28.95%	9.26%	8.51%	0.00%	13.04%	19.35%	10.42%	0.00%	16.00%
GO:0006626	7.89%	0.00%	1.06%	0.00%	4.35%	3.23%	4.17%	0.00%	0.00%
GO:0000749	7.89%	0.00%	1.06%	3.57%	4.35%	0.00%	0.00%	0.00%	0.00%
GO:0016197	7.89%	1.85%	3.19%	0.00%	0.00%	3.23%	0.00%	0.00%	4.00%
GO:0045333	5.26%	1.85%	1.06%	0.00%	0.00%	3.23%	2.08%	0.00%	0.00%
GO:0046467	5.26%	1.85%	1.06%	0.00%	0.00%	0.00%	12.50%	0.00%	0.00%
GO:0000028	5.26%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0006450	5.26%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GO:0007105	2.63%	1.85%	1.06%	3.57%	0.00%	0.00%	2.08%	0.00%	0.00%

De las tablas III.a-i puede notarse que los valores de **Ø** varían mucho entre clases, alcanzando en algunos casos valores hasta del 88% (si no tomamos en cuenta los conjuntos de las clases 7 y 9, que por su alto contenido de genes desconocidos representan un caso especial). Los valores de cobertura más altos se presentaron en las clases 1,2 y 6, que son precisamente las que agrupan un mayor número de genes (tamaño clase 1=1116, tamaño clase 2=1446 y tamaño clase 6=1018). Este fenómeno no es muy sorprendente, ya que las definiciones de estas clases, debido precisamente al gran número de genes que incluyen, contiene términos de GO muy generales, que necesariamente incluyen funciones más específicas, características de otras clases y al momento de determinar si los genes de un subconjunto están contenidos en una definición, los términos generales abarcan mucho más de lo que sería deseable para poder discriminar entre clases. Las definiciones de las clases 3, 4, 5 y 8, que varían en tamaño de 545 a 613 genes, son más excluyentes, obteniéndose en general, valores de **Ø** menores al 75% para los subconjuntos pertenecientes a otras clases. Sólo llama la atención el valor de **Ø** = 87% para el set 8, en relación a la definición de la clase 5 y recíprocamente el valor de **Ø** = 76% para el set 5 cuando se le aplica la definición de la clase 8, lo cual seguramente nos está indicando una gran similitud funcional entre estas dos clases, aunque habría que analizar con más detalle este caso particular para descartar que la similitud se presente debido a que cada uno de ellos cubre primordialmente ramas distintas de una rama más general, que sería la que nos estaría generando la aparente similitud.

Aunque a primera vista esta primera parte de la etapa (**b**) del análisis podría no parecer muy satisfactoria, ya que los genes de las distintas clases están en muchos casos abarcados en un alto porcentaje por las definiciones de otras clases, cuando analizamos la distribución porcentual de los distintos genes en los subconjuntos y los comparamos con la definición de la clase, obtenemos mejores resultados. Con la finalidad de comparar las distribuciones de la definición de la clase con las obtenidas para cada uno de los conjuntos (sets) asignados a otras clases, aplicamos nuevamente la prueba de Wilcoxon de rangos signados. Los resultados de esta prueba estadística para las comparaciones de todos los conjuntos con la definición de cada una de las 9 clases, así como una gráfica comparativa de las distribuciones, se presentan en las figs. 58-66. Para cada una de las gráficas, las primeras cuatro barras corresponden a la definición de la clase y a las tres muestras asignadas a dicha clase, cuyos resultados analizamos anteriormente (**a**), y las demás barras corresponden a los conjuntos asignados a otras clases.

De las nueve gráficas, puede notarse claramente que el patrón de bandas presentado por los conjuntos asignados a la clase (**a**) se asemeja mucho más a la definición de la misma, que el que presentan todos los conjuntos asignados a otras clases (**b**). Este resultado se confirma cuando analizamos los resultados de la prueba estadística aplicada. De las 62 comparaciones que se reportan, sólo en un caso resultó que la distribución de un conjunto no fuera significativamente distinta de la definición de otra clase. Este caso corresponde a la comparación de la distribución del set5 en relación a la definición de la clase 8, lo cual es consistente con los datos que encontramos para **Ø** entre los mismos conjuntos y nos podría estar indicando que los límites entre esas dos clases, al menos desde la perspectiva del nivel de análisis que aquí se hace, no están perfectamente definidos.

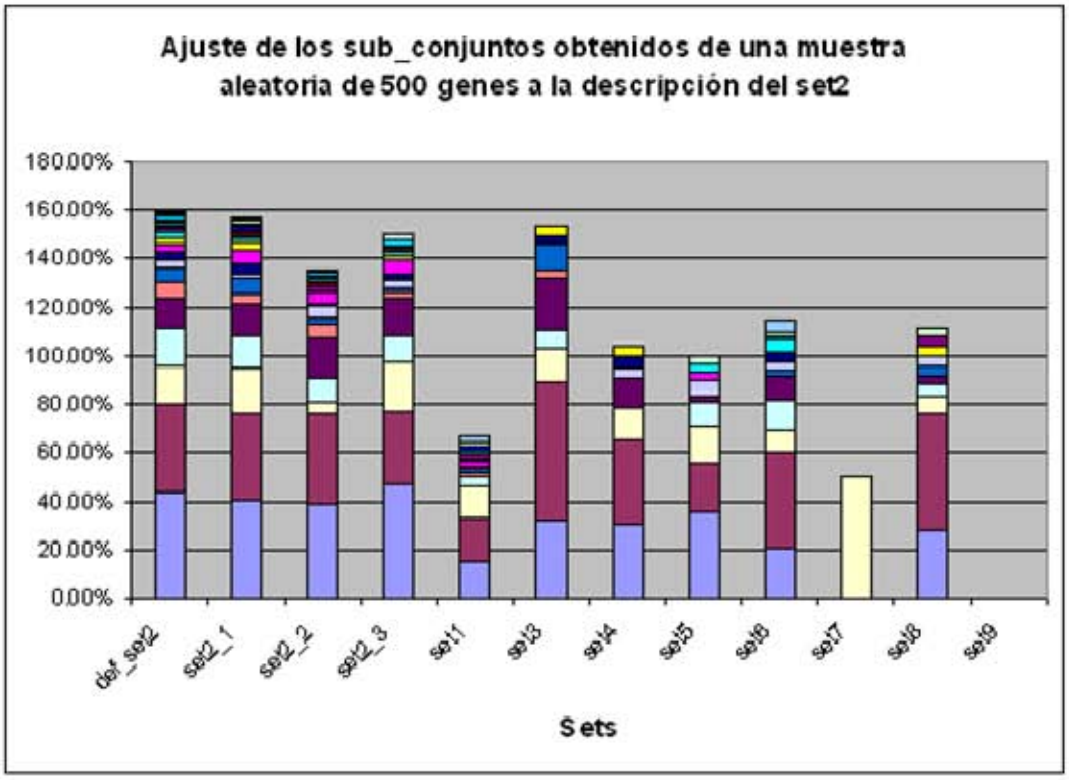


Wilcoxon signed rank test

	Def1_Vs_set2	Def1_Vs_set3	Def1_Vs_Set4	Def1_Vs_set5
P value	0.02	0.0015	0.0031	0.0156
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	643.5 , -302.5	719.0 , -227.0	700.0 , -246.0	652.0 , -294.0
Sum of signed ranks (W)	341	492	454	358
How effective was the pairing?				
rs (Spearman, Approximation)	0.4106	0.14	0.1275	0.4712
P Value	0.0031	0.1853	0.2075	0.0007
P value summary	**	ns	ns	***
Was the pairing significantly effective?	Yes	No	No	Yes

	Def1_Vs_set6	Def1_Vs_set7	Def1_Vs_set8
P value	0.0019	P<0.0001	0.0124
Are medians signif. different? (P < 0.05	Yes	Yes	Yes
Sum of positive, negative ranks	713.0 , -233.0	903.0 , -43.00	631.5 , -271.5
Sum of signed ranks (W)	480	860	360
How effective was the pairing?			
rs (Spearman, Approximation)	0.361	0.112	0.2318
P Value	0.0087	0.2373	0.0673
P value summary	**	ns	ns
Was the pairing significantly effective?	Yes	No	No

. Fig. 58.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 1 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 1 y los subconjuntos asignados a las clases 2-8.

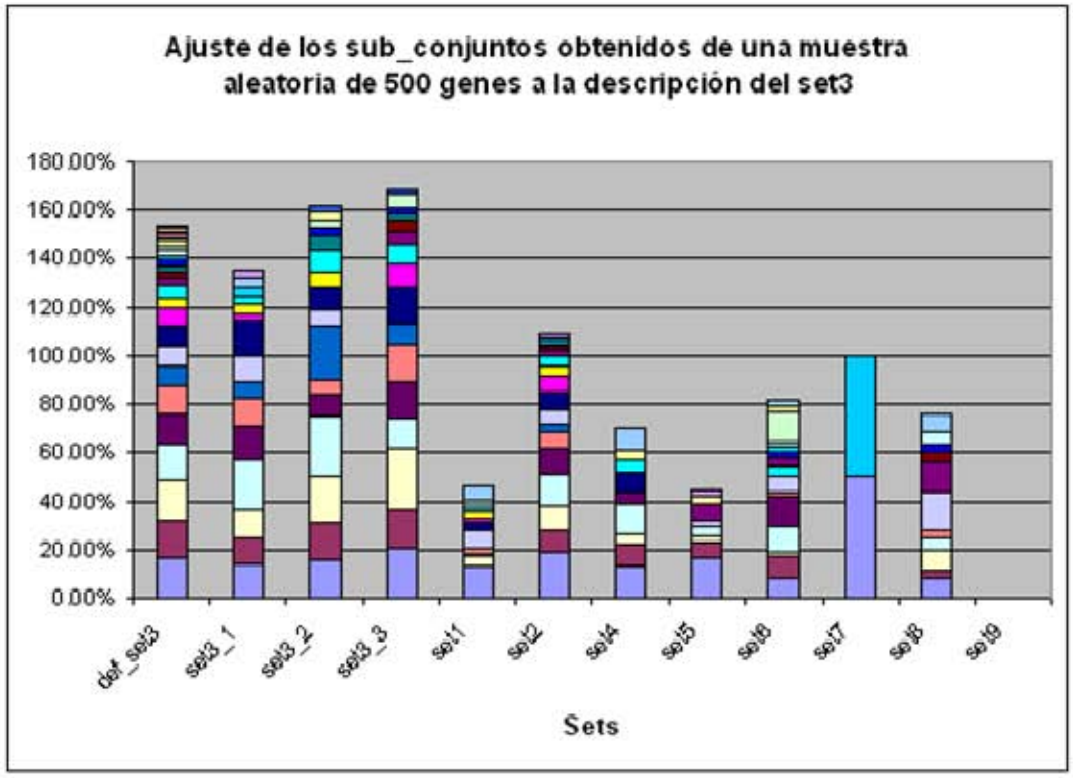


Wilcoxon signed rank test

	Def2_Vs_set1	Def2_Vs_set3	Def2_Vs_Set4	Def2_Vs_set5
P value	0.0012	0.0348	0.0016	0.0024
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	203.0 , -28.00	168.0 , -63.00	201.0 , -30.00	197.0 , -34.00
Sum of signed ranks (W)	175	105	171	163
How effective was the pairing?				
rs (Spearman, Approximation)	0.4917	0.8558	0.6872	0.6685
P Value	0.0118	P<0.0001	0.0003	0.0005
P value summary	*	***	***	***
Was the pairing significantly effective?	Yes	Yes	Yes	Yes

	Def2_Vs_set6	Def2_Vs_set7	Def2_Vs_set8
P value	0.0436	0.0004	0.0216
Are medians signif. different? (P < 0.05	Yes	Yes	Yes
Sum of positive, negative ranks	165.0 , -66.00	212.0 , -19.00	174.0 , -57.00
Sum of signed ranks (W)	99	193	117
How effective was the pairing?			
rs (Spearman, Approximation)	0.6268	0.2955	0.6725
P Value	0.0012	0.0967	0.0004
P value summary	**	ns	***
Was the pairing significantly effective?	Yes	No	Yes

. Fig. 59.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 2 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 2 y los subconjuntos asignados a las clases 1 y 3-8.

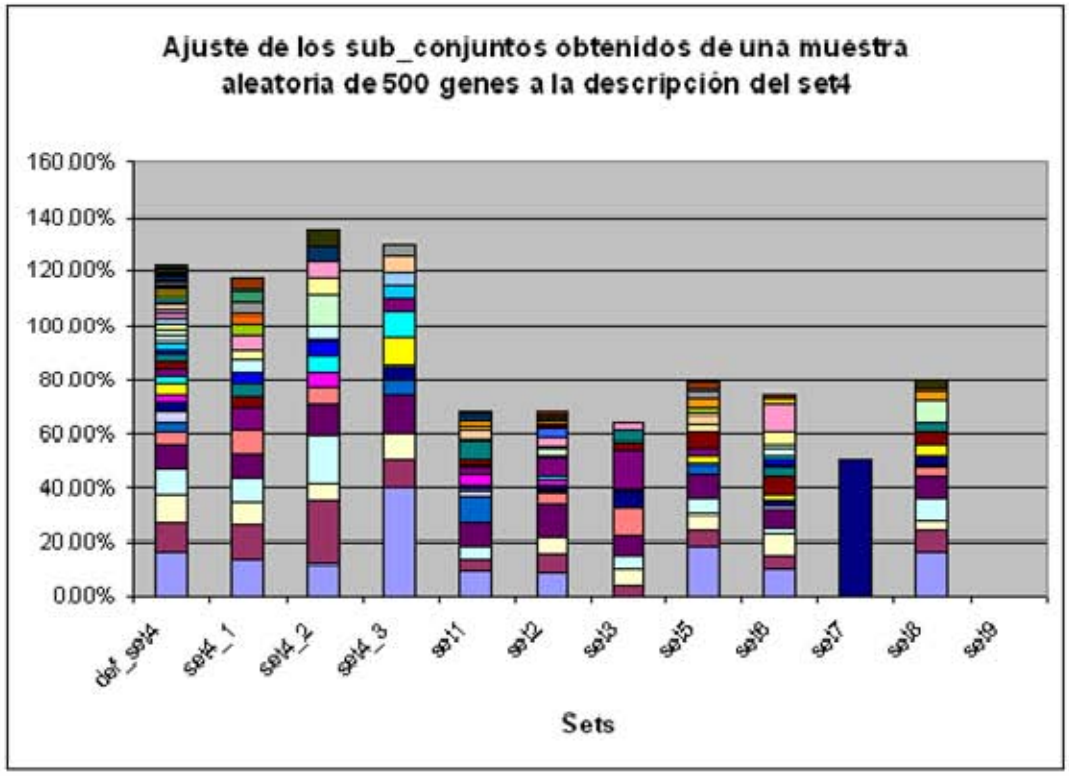


Wilcoxon signed rank test

	Def3_Vs_set1	Def3_Vs_set2	Def3_Vs_Set4	Def3_Vs_set5
P value	P<0.0001	P<0.0001	0.0002	0.0001
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	333.0 , -18.00	328.0 , -23.00	314.0 , -37.00	321.0 , -30.00
Sum of signed ranks (W)	315	305	277	291
How effective was the pairing?				
rs (Spearman, Approximation)	0.395	0.9225	0.5203	0.422
P Value	0.0229	P<0.0001	0.0032	0.0159
P value summary	*	***	**	*
Was the pairing significantly effective?	Yes	Yes	Yes	Yes

	Def3_Vs_set6	Def3_Vs_set7	Def3_Vs_set8
P value	0.0006	0.0008	0.0079
Are medians signif. different? (P < 0.05	Yes	Yes	Yes
Sum of positive, negative ranks	304.0 , -47.00	300.0 , -51.00	271.0 , -80.00
Sum of signed ranks (W)	257	249	191
How effective was the pairing?			
rs (Spearman, Approximation)	0.4827	0.1638	0.4499
P Value	0.0063	0.212	0.0105
P value summary	**	ns	*
Was the pairing significantly effective?	Yes	No	Yes

Fig. 60.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 3 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 3 y los subconjuntos asignados a las clases 1-2 y 4-8.

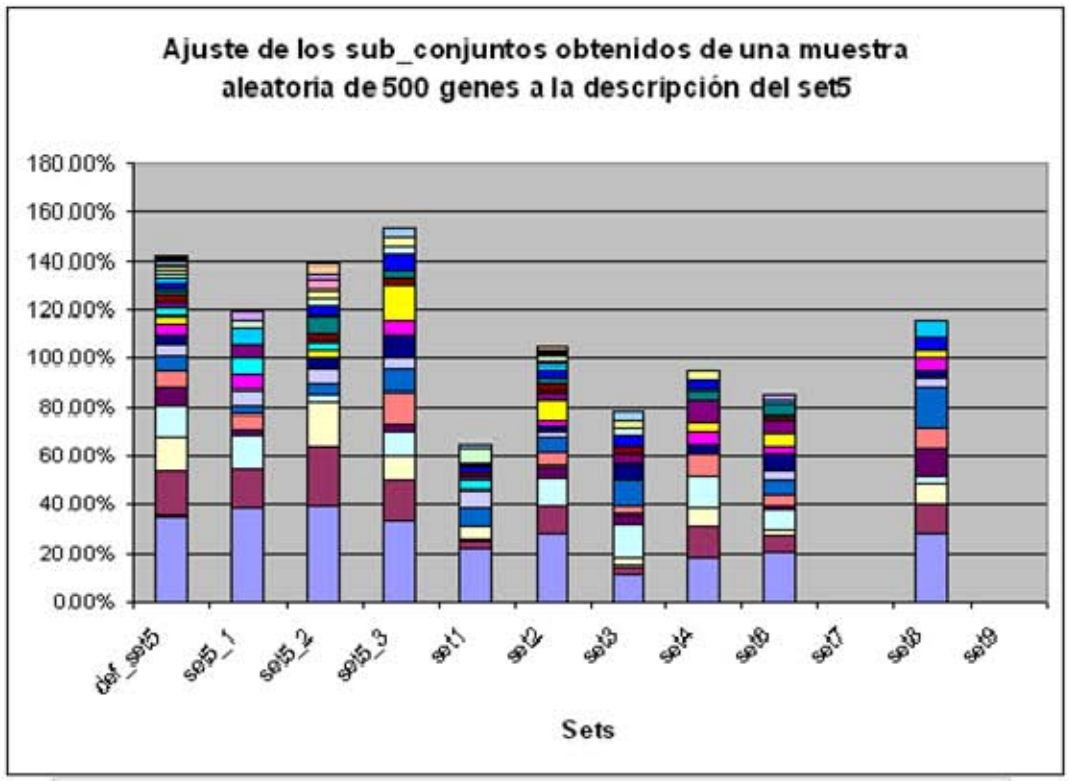


Wilcoxon signed rank test

	Def4_Vs_set1	Def4_Vs_set2	Def4_Vs_Set3	Def4_Vs_set5
P value	0.0004	0.0003	0.0005	0.0041
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	576.0 , -127.0	578.0 , -125.0	572.0 , -131.0	527.0 , -176.0
Sum of signed ranks (W)	449	453	441	351
How effective was the pairing?				
rs (Spearman, Approximation)	0.4598	0.3987	0.5556	0.4179
P Value	0.0021	0.0073	0.0002	0.005
P value summary	**	**	***	**
Was the pairing significantly effective?	Yes	Yes	Yes	Yes

	Def4_Vs_set6	Def4_Vs_set7	Def4_Vs_set8
P value	0.0008	P<0.0001	0.0008
Are medians signif. different? (P < 0.05	Yes	Yes	Yes
Sum of positive, negative ranks	562.0 , -141.0	666.0 , -37.00	562.0 , -141.0
Sum of signed ranks (W)	421	629	421
How effective was the pairing?			
rs (Spearman, Approximation)	0.5056	0.1649	0.5641
P Value	0.0007	0.1646	0.0001
P value summary	***	ns	***
Was the pairing significantly effective?	Yes	No	Yes

Fig. 61.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 4 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 4 y los subconjuntos asignados a las clases 1-3 y 5-8.

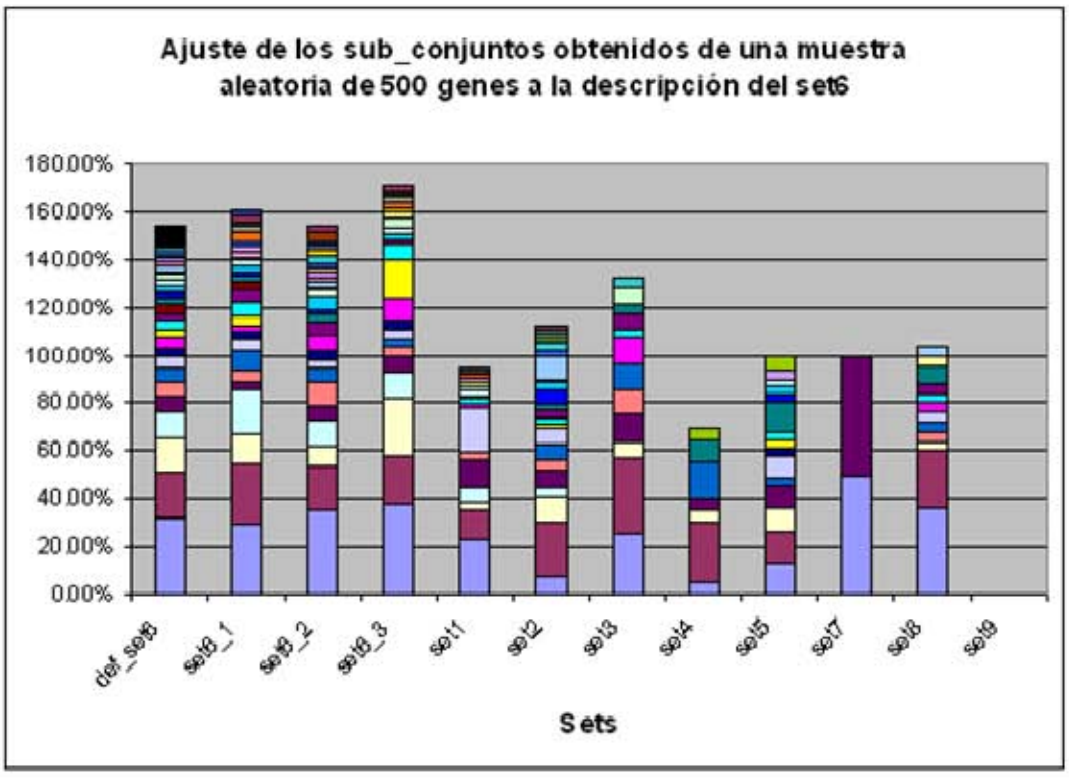


Wilcoxon signed rank test

	Def5_Vs_set1	Def5_Vs_set2	Def5_Vs_Set3
P value	0.0015	0.0126	0.0261
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes
Sum of positive, negative ranks	273.0 , -52.00	246.0 , -79.00	235.0 , -90.00
Sum of signed ranks (W)	221	167	145
How effective was the pairing?			
rs (Spearman, Approximation)	0.3756	0.6734	0.5665
P Value	0.0321	0.0001	0.0016
P value summary	*	***	**
Was the pairing significantly effective?	Yes	Yes	Yes

	Def5_Vs_set4	Def5_Vs_set6	Def5_Vs_set8
P value	0.0261	0.0077	0.0202
Are medians signif. different? (P < 0.05	Yes	Yes	Yes
Sum of positive, negative ranks	235.0 , -90.00	253.0 , -72.00	239.0 , -86.00
Sum of signed ranks (W)	145	181	153
How effective was the pairing?			
rs (Spearman, Approximation)	0.6314	0.7555	0.8071
P Value	0.0004	P<0.0001	P<0.0001
P value summary	***	***	***
Was the pairing significantly effective?	Yes	Yes	Yes

Fig. 62.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 5 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 5 y los subconjuntos asignados a las clases 1-4, 6 y 8.

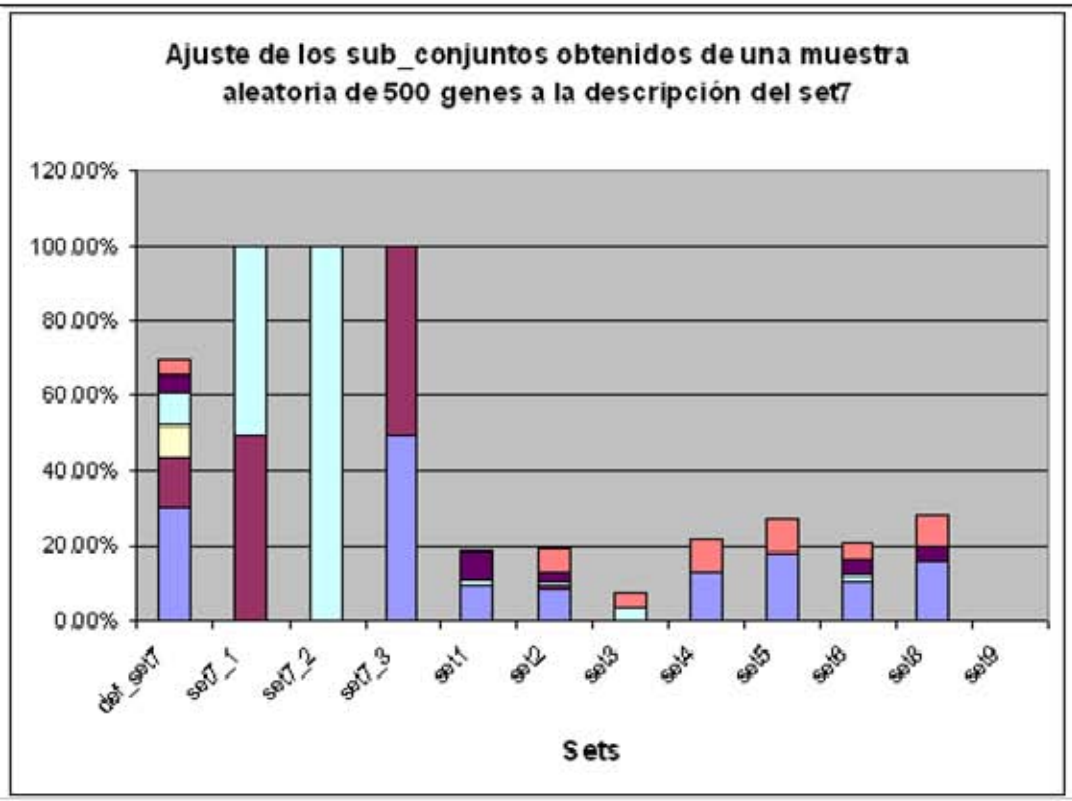


Wilcoxon signed rank test

	Def6_Vs_set1	Def6_Vs_set2	Def6_Vs_Set3	Def6_Vs_set4
P value	P<0.0001	0.0384	0.0351	0.0001
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	706.0 , -114.0	542.0 , -278.0	545.0 , -275.0	681.0 , -139.0
Sum of signed ranks (W)	592	264	270	542
How effective was the pairing?				
rs (Spearman, Approximation)	0.6018	0.6295	0.6415	0.4805
P Value	P<0.0001	P<0.0001	P<0.0001	0.0009
P value summary	***	***	***	***
Was the pairing significantly effective?	Yes	Yes	Yes	Yes

	Def6_Vs_set5	Def6_Vs_set7	Def6_Vs_set8
P value	0.002	P<0.0001	0.0009
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes
Sum of positive, negative ranks	625.0 , -195.0	742.0 , -78.00	644.0 , -176.0
Sum of signed ranks (W)	430	664	468
How effective was the pairing?			
rs (Spearman, Approximation)	0.6257	0.3489	0.6168
P Value	P<0.0001	0.0137	P<0.0001
P value summary	***	*	***
Was the pairing significantly effective?	Yes	Yes	Yes

Fig. 63.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 6 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 6 y los subconjuntos asignados a las clases 1-5 y 7-8.

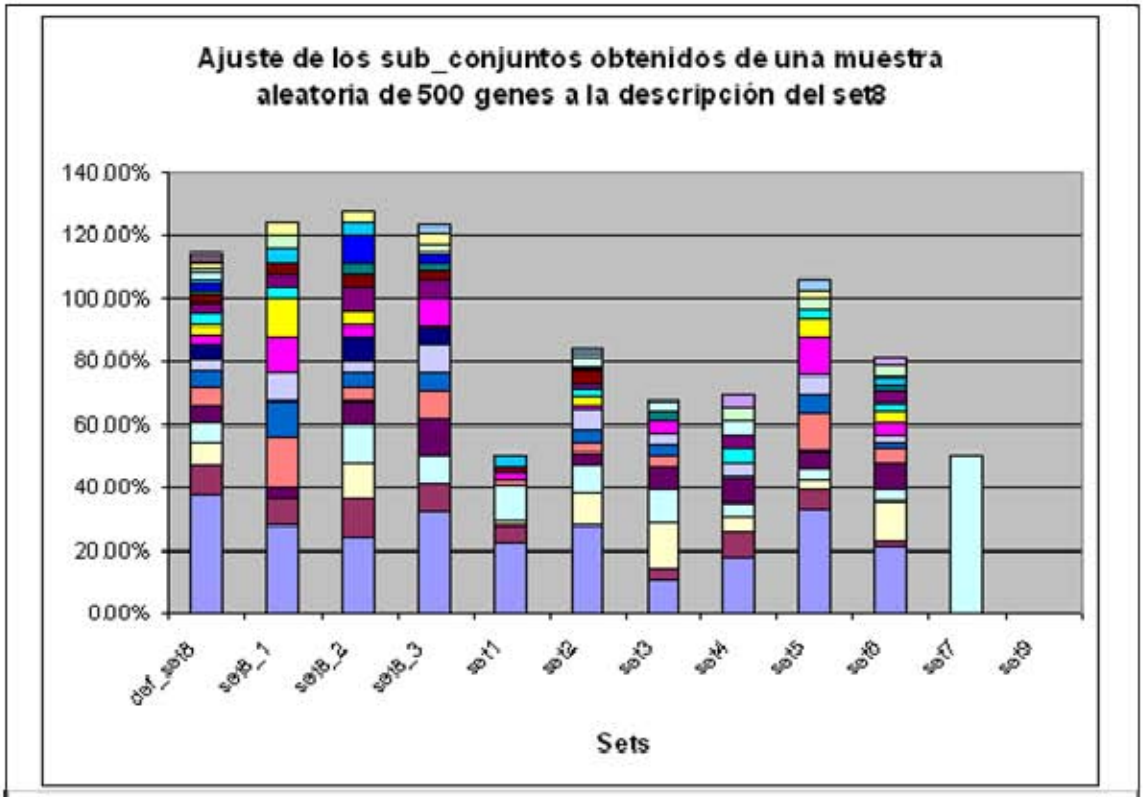


Wilcoxon signed rank test

	Def7_Vs_set1	Def7_Vs_set2	Def7_Vs_Set3	Def7_Vs_set4
P value	0.0313	0.0313	0.0156	0.0313
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	20.00 , -1.000	20.00 , -1.000	21.00 , 0.0000	19.50 , -1.500
Sum of signed ranks (W)	19	19	21	18
How effective was the pairing?				
rs (Spearman, Approximation)	0.2501	0.0448	-0.4264	0.2611
P Value	0.3292	0.4597	0.2097	0.3292
P value summary	ns	ns	ns	ns
Was the pairing significantly effective?	No	No	No	No

	Def7_Vs_set5	Def7_Vs_set6	Def7_Vs_set8
P value	0.0469	0.0156	0.0469
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes
Sum of positive, negative ranks	19.00 , -2.000	21.00 , 0.0000	19.00 , -2.000
Sum of signed ranks (W)	17	21	17
How effective was the pairing?			
rs (Spearman, Approximation)	0.2611	-0.0455	0
P Value	0.3292	0.4597	0.5
P value summary	ns	ns	ns
Was the pairing significantly effective?	No	No	No

Fig. 64.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 7 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 7 y los subconjuntos asignados a las clases 1-6 y 8.

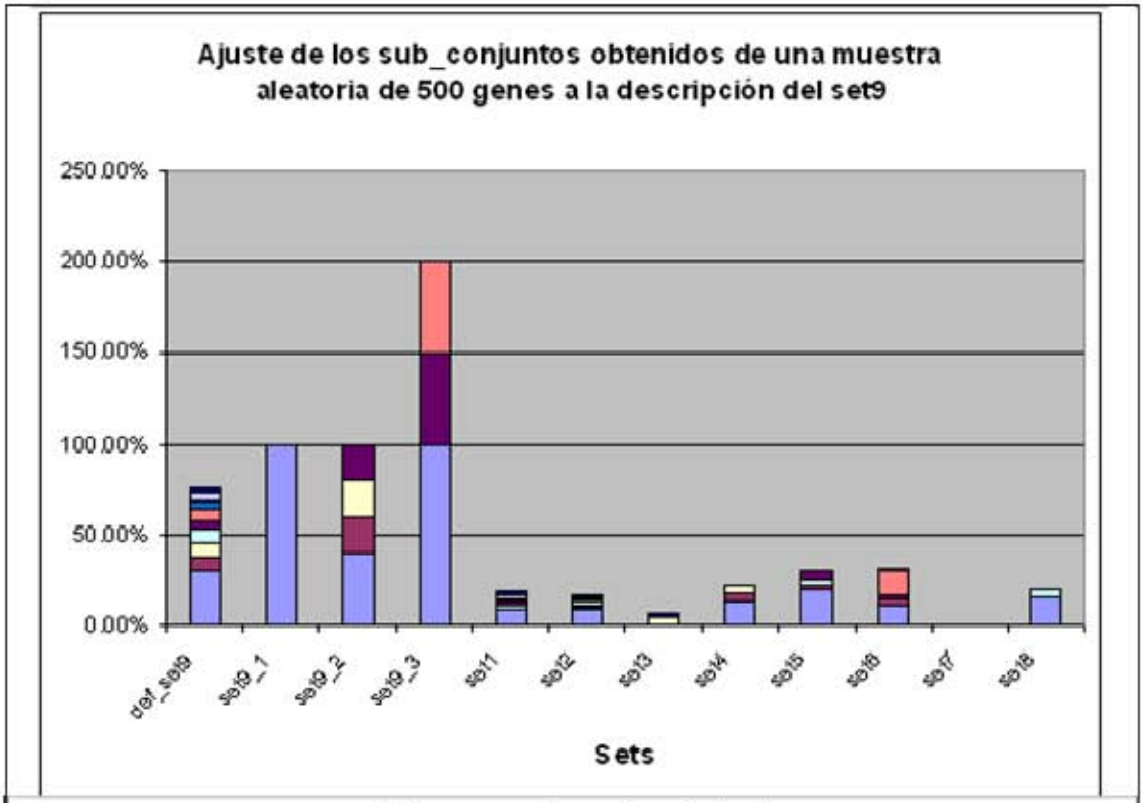


Wilcoxon signed rank test

	Def8_Vs_set1	Def8_Vs_set2	Def8_Vs_Set3	Def8_Vs_set4
P value	0.0005	0.0283	0.0283	0.0158
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	247.0 , -29.00	201.0 , -75.00	201.0 , -75.00	209.0 , -67.00
Sum of signed ranks (W)	218	126	126	142
How effective was the pairing?				
rs (Spearman, Approximation)	0.5804	0.629	0.7441	0.4455
P Value	0.0018	0.0007	P<0.0001	0.0166
P value summary	**	***	***	*
Was the pairing significantly effective?	Yes	Yes	Yes	Yes

	Def8_Vs_set5	Def8_Vs_set6	Def8_Vs_set7
P value	0.1134	0.0325	0.0003
Are medians signif. different? (P < 0.05)	No	Yes	Yes
Sum of positive, negative ranks	178.0 , -98.00	199.0 , -77.00	253.0 , -23.00
Sum of signed ranks (W)	80	122	230
How effective was the pairing?			
rs (Spearman, Approximation)	0.6506	0.6327	0.2577
P Value	0.0004	0.0006	0.1176
P value summary	***	***	ns
Was the pairing significantly effective?	Yes	Yes	No

Fig. 65.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 8 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 8 y los subconjuntos asignados a las clases 1-7.



Wilcoxon signed rank test

	Def9_Vs_set1	Def9_Vs_set2	Def9_Vs_Set3	Def9_Vs_set4
P value	0.002	0.002	0.0039	0.002
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes	Yes
Sum of positive, negative ranks	45.00 , 0.0000	45.00 , 0.0000	44.00 , -1.000	45.00 , 0.0000
Sum of signed ranks (W)	45	45	43	45
How effective was the pairing?				
rs (Spearman, Approximation)	0.0397	0.6345	-0.2203	0.8054
P Value	0.4597	0.0332	0.2845	0.0044
P value summary	ns	*	ns	**
Was the pairing significantly effective?	No	Yes	No	Yes

	Def9_Vs_set5	Def9_Vs_set6	Def9_Vs_set8
P value	0.002	0.0273	0.002
Are medians signif. different? (P < 0.05)	Yes	Yes	Yes
Sum of positive, negative ranks	45.00 , 0.0000	39.00 , -6.000	45.00 , 0.0000
Sum of signed ranks (W)	45	33	45
How effective was the pairing?			
rs (Spearman, Approximation)	0.6741	0.0884	0.6799
P Value	0.0232	0.4106	0.022
P value summary	*	ns	*
Was the pairing significantly effective?	Yes	No	Yes

Fig. 66.- La gráfica muestra la representatividad de los términos de GO correspondientes a la definición de la clase 9 para los subconjuntos asignados a las clases 1-9 a partir de muestras aleatorias de 500 genes. En la tabla se presentan los resultados de la comparación estadística (Prueba de Wilcoxon) de las distribuciones de porcentajes de términos de GO entre la definición de la clase 9 y los subconjuntos asignados a las clases 1-6 y 8.

Los resultados correspondientes a los conjuntos obtenidos a partir de muestras aleatorias de 1000 genes fueron muy semejantes a los que se acaban de presentar y por lo mismo se tomó la decisión de no incluirlos, ya que no aportaban nada nuevo a los resultados.

La conclusión de este análisis sería que las definiciones de las 9 clases nos permiten discriminar entre conjuntos de genes asociados a una clase dada en términos de la composición porcentual de las funciones de los genes que las componen, pero no establecer de manera tajante funciones exclusivas y características de cada una de las clases. Es probable que un análisis más fino, subclasificando los 9 conjuntos iniciales, nos permitiera llegar a un nivel de detalle en la descripción de las funciones asociadas a cada clase mucho más excluyente.

5

DISCUSIÓN Y CONLUSIONES

5.1 Interpretación de los resultados

Los objetivos planteados para este trabajo se cumplieron en su totalidad:

1) Se desarrolló una nueva forma de representación de las secuencias de proteínas que resultó apropiada para su análisis mediante los métodos propuestos.

2) Se obtuvieron clasificaciones consistentes y

3) Se demostró que las clases obtenidas muestran relaciones interesantes con categorías funcionales de las proteínas.

5.1.1 Representación de las secuencias

Pasando a hacer un análisis más puntual de los resultados, podemos decir que el método propuesto para la representación de las secuencias de proteínas es una aportación interesante al campo, ya que aunque existen trabajos previos que han propuesto representaciones vectoriales de las secuencias, para poder luego analizarlas mediante técnicas de computación suave (Ferran E et al., 1994; Han LY et al., 2004), no conocemos ningún trabajo que haya recurrido a una representación en la que se integre información de composición y estructura como la que aquí se propone, siendo además lo suficientemente simple, como para no estar viciada por suposiciones apriorísticas.

Se propusieron cuatro formas alternativas de representar las secuencias de aminoácidos de las proteínas:

a) Ξ_1, con información de la secuencia dispuesta de forma lineal,

b) Ξ_2 con información de la secuencia dispuesta en un plano bidimensional adicional a Ξ_1,

c) Ξ_3 con información de la secuencia dispuesta en una estructura tridimensional más la información de Ξ_2 y

d) Ξ_4 con información de la secuencia dispuesta en un espacio tetradimensional, adicionalmente a la información de Ξ_3.

Se realizaron las clasificaciones y los análisis sobre las cuatro formas de representación. Aunque pretendíamos discriminar entre ellas para seleccionar una forma óptima de representación con fines clasificatorios, esto no fue posible mediante las pruebas realizadas. Sin embargo, podemos concluir que en términos generales, el método de representación propuesto nos permite obtener clasificaciones interesantes y representa una alternativa para la representación de secuencias de aminoácidos, para ser utilizada por otras herramientas de análisis con este u otros fines.

5.1.2 Métodos de clasificación

Se trabajó con dos métodos distintos para la clasificación de las secuencias: fuzzy-c means y mapas autoorganizados. Los algoritmos utilizados por cada uno de ellos son muy distintos y no puede dejar de llamar la atención el hecho de que las clasificaciones obtenidas para cada una de nuestras representaciones de las secuencias (Ξ_1- Ξ_4) hayan sido muy semejantes (fig.23).

Este hecho podría tener dos interpretaciones: 1) Nuestra representación de las secuencias rescata diferencias triviales entre las cadenas de proteínas, mismas que pueden ser detectadas por cualquier método clasificatorio. De ser este el caso, los análisis que hicimos comparando longitud de la cadena y frecuencia de aminoácidos habrían mostrado diferencias significativas entre los grupos, lo cual no sucedió (figs.25-32). 2) Nuestra representación de las secuencias, por el contrario, tiene la capacidad de rescatar características relevantes, asociadas tanto a la composición como a la estructura de las secuencias de aminoácidos, pero no detectables de manera trivial, lo que permite una delimitación "natural" de los grupos e independientemente del método utilizado, éstos emergen como entidades distintas. Los análisis realizados sobre los conjuntos obtenidos apuntan fuertemente en esta dirección, ya que las relaciones funcionales que se encontraron entre los genes pertenecientes a cada una de las clases no pudieron aparecer de manera azarosa.

5.1.3 Análisis de las clases obtenidas

Ya en la sección de resultados (4.3.2) hicimos algunas puntualizaciones respecto a las clasificaciones obtenidas, pero en esta sección intentaré resumirlas en un análisis más general.

1) La primera y más importante conclusión a la que pudimos llegar es que las clases obtenidas mediante nuestro método presentan una alta correlación con categorías funcionales previamente descritas en la base de datos Gene Ontology, al compararlas con conjuntos generados de manera aleatoria (figs.33-51). Este hecho, por un lado confirma el principio básico del plegamiento de las proteínas, en el sentido de que la estructura primaria determina la forma y función de las proteínas, y por otro lado, nos permite reafirmar nuestro método como una forma no sesgada de clasificación de las proteínas, que nos lleva a grupos distinguibles entre sí y con significado biológico.

2) Por otro lado el análisis de los términos de GO que resultaron significativos en nuestros conjuntos (figs.53-54) nos permite afirmar, que las proteínas asociadas a los mismos se concentran en ramas del árbol de GO, lo que nos indica relaciones funcionales y probablemente evolutivas entre ellas, en contraste con los patrones dispersos que se presentaron para los conjuntos aleatorios.

3) Por último, el análisis de los conjuntos mostró que las proteínas de función desconocida no se distribuyen de manera uniforme entre nuestras clases (figs.55-56), habiendo una mayor proporción de proteínas desconocidas asociadas a conjuntos pequeños, generalmente de cadena corta y relacionados con proteínas con función muy especializada, como homeostasis de metales, transporte mediado por vesículas y transporte de metales, mientras que los conjuntos asociados a funciones más generales del metabolismo, reproducción y ciclo celular, tienden a ser conjuntos más grandes, con proteínas de cadena más larga y con una menor proporción de genes de función desconocida. Este hecho podría interpretarse como que las proteínas que participan en procesos celulares y metabólicos fundamentales para la vida son en su mayoría ya conocidas y ampliamente estudiadas por bioquímicos, fisiólogos y biólogos celulares, quedando sólo pocas de éstas por caracterizar, mientras que aquéllas, responsables de procesos más especializados, relacionados con funciones muy específicas de regulación, homeostasis o transporte de pequeñas moléculas, aún no han sido suficientemente caracterizadas y faltan muchas por describir y analizar.

Un análisis mucho más profundo de los conjuntos, así como la comparación con otras bases de datos y otras clasificaciones, nos permitiría llegar a conclusiones más contundentes con respecto a nuestros conjuntos y ayudaría a distinguir la potencialidad clasificatoria de nuestras cuatro representaciones para los vectores de características (Ξ_1- Ξ_4).

5.1.4 Extensión del método

Como mencionamos en la sección **4.2.3**, una de las razones por las cuales decidimos utilizar redes de Kohonen para la clasificación de las secuencias de proteínas fue la particularidad de este método de permitirnos hacer posteriores subclasificaciones de las secuencias para obtener un mayor número de conjuntos a partir de las nueve clases iniciales. Esta característica del método la ejemplificamos en la fig.22, en donde se puede ver la subdivisión de las clases originales en varios subconjuntos etiquetados.

Este paso ya no fue posible realizarlo, debido a la cantidad de análisis adicionales que involucra, pero sería una continuación natural de este trabajo, que permitiría llegar a estructuras clasificatorias más complejas, tipo árbol, en las que podría hacerse un análisis más detallado de la forma en que las proteínas se distribuyen dentro de cada una de nuestras clases iniciales.

5.2 Aportaciones del método

Aunque los resultados presentados en este trabajo no son aún concluyentes, si podemos decir que se trata de una aportación interesante para el área de clasificación de proteínas desde dos perspectivas distintas.

5.2.1 Aportaciones relacionadas con el enfoque

Como se menciona en repetidas ocasiones en la introducción, el campo de la clasificación de proteínas ha estado dominado durante la última década por enfoques muy reduccionistas y apriorísticos, en los que los alineamientos de las proteínas y más concretamente de pequeñas regiones de éstas son usados como el punto de partida de las clasificaciones, que luego son acomodadas de acuerdo con la supervisión de los expertos en las categorías más adecuadas. Si bien con estos métodos se han podido descubrir aspectos importantes relacionados con la evolución y la función de las proteínas, parece que han llegado a un punto en el que estos avances son ya muy limitados y no parece haber aportes sustanciales en lo que a la naturaleza y clasificación de las proteínas se refiere. Se obtienen buenas clasificaciones para un 50 o 60% de las cadenas, pero queda un 40% que se muestra difícil de acomodar por casi todos los métodos propuestos (Yona G et al., 1999). Como mencioné anteriormente, el campo parece estar atrapado en su propio paradigma reduccionista.

En contraparte, el método que nosotros proponemos parte de un nuevo enfoque mucho más holista y menos dirigido. Pretendemos considerar a las proteínas ya no como fragmentos de regiones funcionales, sino como entidades completas y tratamos de permitir que si existen patrones o rasgos característicos de cada una de ellas, puedan emerger de manera automática, sin la intervención de un experto que tenga que identificarlos y extraerlos manualmente.

Existen muchos ejemplos en la biología en los que el enfoque reduccionista ha conseguido hacer aportes hasta un cierto nivel y luego ha sido necesaria una aproximación desde otro enfoque mucho más holista para poder conseguir nuevos

avances (Goodenough, 2006; Mayr E, 2006; Van Regenmortel, 2004) y nosotros pensamos que éste podría ser un caso semejante.

5.2.3 Aportaciones asociadas al método

Por otro lado, y como ya se mencionó en los párrafos anteriores, desde el punto de vista metodológico también este trabajo tienen varias aportaciones importantes. Una relacionada con la representación de las secuencias, otra con la utilización de técnicas de computación suave al problema de la clasificación de proteínas, que no había sido usado en la forma propuesta por nosotros; y por último, en la utilización de la base de datos Gene Ontology para la validación de las clasificaciones, y más específicamente en el desarrollo de una herramienta que permite generar mapas funcionales a partir de conjuntos de genes, que si bien no se presenta completa en este trabajo porque es un proyecto independiente (Coello G, 2006), aquí se generaron las bases para su desarrollo.

5.3 Perspectivas

5.3.1 Reconocimiento de patrones y compresión

Como se menciona en los objetivos de este trabajo, una idea importante sobre la que se sustenta la metodología propuesta es el hecho de que existan patrones no triviales en las cadenas de proteínas, que puedan subyacer a la estructura tridimensional y función de las mismas. La hipótesis de trabajo asume que en caso de existir, estos patrones podrían emerger de manera natural mediante nuestro proceso de clasificación y de ser así, un análisis posterior de las cadenas en cada uno de nuestros conjuntos mediante algún algoritmo de reconocimiento de patrones, como el propuesto por Kuri, Galavíz y Herrera (Kuri-Morales AF et al., 2006), podría recuperar esos patrones y tratar de identificar aquéllos que sean exclusivos o característicos de cada una de nuestras clases. Este es un análisis muy interesante, que constituiría un proyecto de trabajo en sí mismo y que podría así mismo aportar elementos para encontrar un método de compresión de las cadenas de aminoácidos.

Cabe señalar que en este punto se hace referencia, obviamente, a patrones no triviales, es decir patrones discontinuos y tal vez de gran longitud que no podrían ser identificados fácilmente por el ojo humano. Es en este sentido que la aportación sería novedosa, ya que no nos referimos a los tradicionales motifs, que durante años han sido descritos e identificados por los investigadores en el área y que corresponden a regiones estructurales y funcionales perfectamente caracterizadas, sino a combinaciones no-lineales de aminoácidos, que podrían no tener necesariamente una estructura tridimensional identificable, pero que podrían ubicarse en el centro de la caracterización de ciertos grupos de proteínas.

5.3.2 Análisis adicionales. Comparación con otros métodos

Habiendo mostrado, en la sección 4.4 de los resultados que nuestro clasificador es capaz de asignar función biológica a un conjunto de proteínas, sería interesante continuar con este ejercicio, subclasificando los conjuntos obtenidos mediante redes de Kohonen y generar así nuevas definiciones más compactas de las clases así obtenidas para conseguir asignar de manera más precisa y exclusiva función biológica a nuevas proteínas.

Para que los resultados de este trabajo tengan una mayor validez sería muy importante hacer una comparación de nuestros resultados con los que se han

obtenido por otros métodos de clasificación. Sin embargo, este paso resulta bastante complejo, ya que no todos los trabajos presentan el listado de genes asociados a sus clasificaciones, y sería necesario realizar la comparación contra una clasificación equivalente, es decir sobre la totalidad de los genes de *S. cerevisiae* y no hemos encontrado trabajos con esas características.

Hicimos un trabajo exploratorio contra la clasificación que acompaña a la base de datos iProClass (Wu CH et al., 2004) en el sitio de PIR (Protein Information Resources http://pir.georgetown.edu/home.shtml), pero encontramos que la cantidad de familias reportadas es mucho mayor que las nuestras, por lo que análisis resultó bastante complejo y aún no tenemos ningún resultado concluyente. Es posible que cuando tengamos una clasificación más detallada, con un mayor número de conjuntos, este paso sea más viable.

APÉNDICE 1

APUNTES SOBRE LOS ESTADÍSTICOS UTILIZADOS

Para responder una pregunta desde el punto de vista estadístico, la pregunta debe ser traducida en forma de una hipótesis, una afirmación que pueda ser sujeta a prueba. Dependiendo del resultado de la prueba, la hipótesis se acepta o se rechaza.

La hipótesis que se pondrá a prueba se conoce como la hipótesis nula ($\mathbf{H_0}$) y para cada hipótesis nula debe existir una hipótesis alternativa ($\mathbf{H_A}$). La hipótesis nula tiene mayor prioridad y no se rechazará a menos de que exista fuerte evidencia en su contra. Si una de las dos hipótesis es más simple se le dará la prioridad, de forma que la teoría más complicada no se adoptará a menos de que haya suficiente evidencia en contra de la más simple. En general es más simple proponer que no existe diferencia entre dos conjuntos de resultados que decir que sí existe una diferencia.

El resultado de someter una hipótesis a prueba será "rechazar H_0" o "no rechazar H_0". Si la conclusión es "no rechazar H_0" esto no significa necesariamente que la hipótesis nula sea cierta, sólo significa que no hay suficiente evidencia en contra de H_0 que favorezca a H_A. Rechazar la hipótesis nula significa que la hipótesis alternativa podría ser cierta.

Para decidir si aceptar o rechazar la hipótesis nula se usa el nivel de significancia (α) del resultado ($\alpha \leq 0.05$ ó $\alpha \leq 0.01$). Esto nos permite establecer si existe una "diferencia significativa" entre las poblaciones, es decir, si las diferencias observadas se deben solamente al azar.

Así, el procedimiento para probar una hipótesis sería:

1. Definir H_0 y H_A, con base en lo explicado anteriormente.
2. Escoger el valor de α.
3. Calcular el valor de la prueba estadística.
4. Comparar el valor calculado con una tabla de valores críticos del estadístico aplicado.
5. Si el valor calculado es menor que el valor crítico de la tabla, se acepta la hipótesis nula (H_0). Si el valor es mayor o igual al valor crítico de la tabla, se rechaza la hipótesis nula y se acepta la hipótesis alternativa (H_A).

Aquí es importante notar que una prueba de significancia nunca prueba la hipótesis nula, solamente no nos permite rechazarla. Un valor muy pequeño de P ($P \leq 0.001$) no significa necesariamente una diferencia muy grande, sino que las diferencias observadas son altamente improbables a la luz de la hipótesis nula.

Estadística no-paramétrica

Cuando se analizan datos medidos por una variable cuantitativa continua, las pruebas estadísticas de estimación y contraste frecuentemente empleadas se basan en suponer que se ha obtenido una muestra aleatoria de una distribución de probabilidad de tipo normal o de Gauss. Pero en muchas ocasiones esta suposición no resulta válida, y en otras la sospecha de que no sea adecuada no resulta fácil de comprobar, por tratarse de muestras pequeñas. En estos casos disponemos de dos posibles mecanismos: los datos se pueden transformar de tal manera que sigan una distribución normal, o bien se puede acudir a pruebas estadísticas que no se basan en ninguna suposición en cuanto a la distribución de probabilidad a partir de la que fueron obtenidos los datos, y por ello se denominan **pruebas no paramétricas** (*distribution*

free), mientras que las pruebas que suponen una distribución de probabilidad determinada para los datos se denominan pruebas paramétricas.

Dentro de las pruebas paramétricas, las más habituales se basan en la distribución de probabilidad normal, y al estimar los parámetros del modelo se supone que los datos constituyen una muestra aleatoria de esa distribución, por lo que la elección del estimador y el cálculo de la precisión de la estimación, elementos básicos para construir intervalos de confianza y contrastar hipótesis, dependen del modelo probabilístico supuesto.

Prueba de χ_i cuadrado

La prueba de chi-cuadrado es una prueba no paramétrica que mide la discrepancia entre una distribución observada y otra teórica (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar. También se utiliza para probar la independencia de dos muestras entre sí, mediante la presentación de los datos en tablas de contingencia.

La fórmula para obtener el estadístico es la siguiente:

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Los grados de libertad nos vienen dados por :

gl= (*r*-1)(*k*-1). Donde *r* es el número de renglones y *k* el de columnas o muestras.

- Criterio de decisión:

Se acepta $\mathbf{H_0}$ cuando el valor de χ^2 obtenido es menor al valor crítico reportado en las tablas para los grados de libertad correspondientes y el nivel de significancia elegido. En caso contrario se rechaza $\mathbf{H_0}$ y se acepta $\mathbf{H_A}$.

Cuanto más se aproxima a cero el valor de chi-cuadrado, más ajustadas están ambas distribuciones.

Cualquier prueba de significancia estadística aplicada de manera apropiada nos permite obtener el grado de confianza que podemos tener al aceptar o rechazar una hipótesis (H_0). Típicamente la hipótesis que se pone a prueba con chi cuadrada es si dos muestras son suficientemente diferentes entre sí en alguna característica o aspecto de su comportamiento de forma que podamos generalizar a partir de nuestras muestras que las poblaciones de las cuales fueron tomadas dichas muestras son también diferentes en su comportamiento o características.

Esta prueba la aplicamos para comparar las distribuciones de genes idénticos asociados a nuestros dos métodos de clasificación: mapas autoorganizados y fuzzy c-means en la sección **4.2.3** de la tesis.

Z- Score

Z score es básicamente una manera de convertir un score (valor) a un número de desviaciones estándar con respecto a la media de la población. El z score de un dato indica la magnitud y la dirección en la que el valor se aleja de la media, expresado en número de desviaciones estándar. Las matemáticas subyacentes a la transformación en z-scores son tales que si todos los datos de la distribución son convertidos a z-score, los valores transformados tendrán necesariamente una media de cero y una desviación estándar de uno.

Los z-scores son también llamados valores-estándar. Esta transformación es particularmente útil cuando queremos comparar las posiciones relativas de valores provenientes de distribuciones con diferentes medias y/o diferentes desviaciones estándar.

La fórmula que debemos aplicar para realizar la conversión de un valor **x** a un valor estándar o z-score (Z) es:

$$Z = \frac{x - \mu}{\sigma}$$

Donde x es el valor que queremos convertir, μ es la media de la población y σ la desviación estándar. De esta forma, Z nos indica el número de desviaciones estándar que nuestro valor se aleja de la media de la población.

Este procedimiento lo utilizamos para comparar las longitudes promedio de las cadenas de aminoácidos de los genes asociados a cada uno de los conjuntos obtenidos mediante nuestro clasificador, en la sección **4.3.1** de la tesis. Los valores reportados en las tablas de las figuras 24, 26, 28 y 30 nos indican a cuántas desviaciones estándar se ubican las medias de cada conjunto, con respecto a los demás conjuntos obtenidos.

Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov se usa para determinar si dos conjuntos de datos difieren significativamente entre sí. Este contraste, que es válido únicamente para variables continuas, compara la función de distribución teórica con la observada, y calcula un valor de discrepancia, representado habitualmente como *D*, que corresponde a la discrepancia máxima en valor absoluto entre la distribución observada y la distribución teórica. Proporciona asimismo un valor de probabilidad *P*, que corresponde a la probabilidad de obtener una distribución que discrepe tanto como la observada si verdaderamente se hubiera obtenido una muestra aleatoria, de tamaño *n*, de la misma población. Si esa probabilidad es grande no habrá por tanto razones estadísticas para suponer que nuestros datos no proceden de la misma distribución, mientras que si es muy pequeña, no será aceptable suponer ese modelo probabilístico para los datos.

Para comparar una muestra de datos consistente de *N* eventos cuya distribución acumulada es $S_N(x)$ con una función hipotética, cuya distribución acumulada es $F(x)$, el valor de D_N se calcula como:

$$D_N = \max\left|S_N(x) - F(x)\right| \quad \text{Para todas las x.}$$

Esta prueba la utilizamos en la sección **4.3.1.** de la tesis para comparar las distribuciones de frecuencias de aminoácidos entre conjuntos.

Prueba de Mann Whitney

La prueba Mann Whitney es un método no paramétrico aplicado a dos muestras independientes cuyos datos han sido medidos en una escala de nivel ordinal (los valores pueden ser ordenados de mayor a menor). La prueba calcula el llamado estadístico *U*. Esta prueba es una alternativa de la prueba de *t* para grupos independientes cuando las suposiciones de normalidad o varianzas iguales no se cumplen. Al igual que en muchas otras pruebas no paramétricas, Mann Whitney utiliza los rangos de los datos en lugar de los valores originales para calcular el estadístico. Se requiere tener dos muestras de tamaño n_1 y n_2 respectivamente. Se ordenan juntas todas las observaciones de ambas muestras y a cada observación se le asigna un rango que va de 1 a N (donde $N=n_1+n_2$) y luego se suman los valores de los rangos para cada muestra (R_1 y R_2). El valor de U_1 puede entonces calcularse con la siguiente fórmula (equivalente para U_2):

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

El valor del estadístico U será el menor entre U_1 y U_2 y deberá entonces compararse con los valores críticos reportados en las tablas. El resultado más importante es el P-value, que nos permite responder a la siguiente pregunta: Si las poblaciones tienen en realidad la misma mediana, ¿cuál es la probabilidad de que un muestreo aleatorio resulte en valores de la mediana tan distintos como los que observamos? Si el P-value es muy pequeño, entonces será poco probable que las diferencias observadas puedan atribuirse exclusivamente al azar. En este caso puede concluirse que las poblaciones tienen diferentes medianas. Si por el contrario, el P-value es grande, significa que no hay suficiente evidencia para concluir que las medianas de ambas poblaciones sean distintas, lo que no es equivalente a decir que las medianas sean iguales.

En la sección **4.3.2** de la tesis aplicamos esta prueba para comparar los valores del índice *Q* entre nuestros conjuntos y conjuntos aleatorios de tamaño semejante. Escogimos esta prueba porque los datos no pasaron la prueba de normalidad y necesitábamos una prueba no pareada debido a que los valores del índice *Q* de los distintos conjuntos no se correspondían entre sí. También se utilizó en la sección **4.3.3** para comparar los valores promedio del índice *Q* obtenidos para todos los conjuntos generados a partir de las matrices $\boldsymbol{\Xi_1}$ a $\boldsymbol{\Xi_4}$.

Prueba de Kruskal-Wallis

La prueba de Kruskal-Wallis es un método no paramétrico que permite comparar tres o más grupos de datos no pareados. La hipótesis que se pone a prueba es si las medianas de al menos dos de las muestras son suficientemente distintas entre sí, como para poder afirmar que proceden de poblaciones distintas. Es la alternativa no paramétrica a la prueba ANOVA (análisis de varianza) y puede considerarse como una generalización de la prueba de Mann Whitney. Al igual que en esta última, todos los datos son ordenados y se les asigna un rango según su posición. El valor del estadístico de Kruskal-Wallis para k muestras independientes se obtiene mediante la siguiente ecuación:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Donde R_i y n_i corresponden a la suma de los rangos y al número de observaciones o datos en la muestra *i* respectivamente.

Un valor grande de *H* corresponde a diferencias importantes entre las sumas de los rangos de las distintas muestras y estará asociado a un P-value pequeño, que nos indica una probabilidad muy baja de que las diferencias observadas entre las muestras puedan deberse al azar, por lo que no podemos afirmar que todas las muestras analizadas provengan de una población con la misma mediana. Esto no significa que existan diferencias entre todas las muestras, sino que al menos dos de ellas presentan grandes diferencias entre sí. Por el contrario, valores pequeños de *H* se asocian con P-values grandes y nos indican una muy alta probabilidad de que las diferencias observadas en las medianas entre muestras puedan atribuirse al azar.

La prueba de Kruskal-Wallis la utilizamos en la sección **4.3.2** de la tesis para comparar los valores del índice *Q* entre todos los conjuntos de tamaño semejante y en la sección **4.3.3** para comparar los valores del índice *Q* obtenidos para todas las clases generadas a partir de las matrices Ξ_1-Ξ_4.

Prueba pareada de rangos de Wilcoxon

Es la alternativa no paramétrica a la prueba de *t* para datos pareados y se utiliza para determinar las diferencias en la mediana entre las muestras. Al igual que en las dos pruebas anteriores, el cálculo del estadístico se hace utilizando rangos en lugar de los valores originales. La forma de calcularlo es obteniendo la diferencia absoluta entre cada par de observaciones, ordenando luego estos valores por orden de magnitud y asignándoles un rango. Luego se calcula la suma de rangos para los valores positivos y negativos y se obtiene la diferencia entre ambos. Este será el estadístico de Wilcoxon (*W*). Por último, se utilizan las tablas de valores críticos para determinar la probabilidad de obtener la *W* encontrada.

Cuanto mayor sea la diferencia entre las sumas de rangos positivos y negativos, menor será el P-value y esto se debe interpretar como que existe una probabilidad muy baja de que las diferencias observadas entre las medianas de las muestras puedan atribuirse al azar y por lo mismo, podemos concluir que no proceden de la misma población. Por el contrario, P-values grandes nos indican una alta probabilidad de que las diferencias observadas entre las medianas puedan deberse al azar y por lo mismo, no tenemos elementos suficientes para afirmar que las muestras provengan de poblaciones distintas.

Utilizamos esta prueba en la sección **4.4** de la tesis para comparar las distribuciones de frecuencias de genes asociados a cada categoría de GO para conjuntos de genes asignados a una clase, con respecto a una definición dada.

APÉNDICE 2

Tabla A2-I .- Muestra los valores de frecuencia (Media y Desviación Estándar) para cada aminoácido en todos los conjuntos generados a partir del vector de características de una dimensión.

AA´s	Global		Set1		Set2		Set3		Set4		Set5		Set6		Set7		Set8		Set9	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	0.0562	0.026	0.0676	0.0216	0.0405	0.0267	0.0799	0.0514	0.0453	0.0122	0.0449	0.0153	0.0494	0.0158	0.0716	0.0179	0.0541	0.0215	0.0752	0.0356
C	0.016	0.0153	0.0209	0.0137	0.0328	0.0232	0.0192	0.0285	0.0142	0.0097	0.0164	0.0097	0.0114	0.0105	0.0137	0.009	0.008	0.0099	0.012	0.0194
D	0.0528	0.0225	0.0362	0.014	0.0237	0.0187	0.0371	0.0221	0.0635	0.0144	0.0558	0.017	0.0557	0.0178	0.0597	0.0144	0.0721	0.0254	0.038	0.0238
E	0.0603	0.0275	0.0396	0.0147	0.0245	0.0196	0.0393	0.0259	0.0697	0.016	0.0616	0.0178	0.0542	0.0169	0.0658	0.0148	0.1062	0.0267	0.0555	0.0288
F	0.0482	0.0244	0.0653	0.0199	0.0849	0.0399	0.0438	0.03	0.0457	0.0135	0.0492	0.0168	0.0366	0.0134	0.0416	0.0125	0.0327	0.0156	0.0375	0.0229
G	0.0505	0.0236	0.0658	0.0224	0.0331	0.0244	0.0557	0.0422	0.0457	0.0129	0.0402	0.0155	0.0445	0.0183	0.0688	0.016	0.0428	0.018	0.0585	0.0294
H	0.0224	0.014	0.0208	0.0129	0.031	0.0277	0.0162	0.0201	0.0228	0.0086	0.0217	0.0097	0.0251	0.0106	0.0221	0.0091	0.0176	0.0108	0.0219	0.0174
I	0.0651	0.0219	0.0737	0.0198	0.0796	0.039	0.0497	0.0238	0.0679	0.0139	0.0725	0.0176	0.0543	0.0148	0.0646	0.0146	0.0552	0.0179	0.0513	0.0246
K	0.0724	0.0297	0.0489	0.0171	0.0483	0.03	0.0408	0.0268	0.079	0.0158	0.0737	0.0204	0.0669	0.02	0.072	0.0162	0.1018	0.0285	0.1186	0.0429
L	0.0971	0.0289	0.103	0.0194	0.1337	0.0437	0.081	0.0383	0.092	0.0126	0.1241	0.0142	0.0791	0.0188	0.0846	0.0137	0.0903	0.0251	0.0801	0.0307
M	0.0229	0.0125	0.0267	0.0107	0.0316	0.024	0.023	0.0161	0.0206	0.0072	0.0224	0.0099	0.021	0.0082	0.0214	0.0092	0.0216	0.0115	0.0235	0.0166
N	0.0567	0.0232	0.045	0.015	0.0444	0.0272	0.0468	0.0365	0.0648	0.0148	0.0599	0.0184	0.0823	0.0234	0.0478	0.0131	0.0571	0.0197	0.046	0.0244
P	0.0438	0.0195	0.0427	0.0158	0.0463	0.0298	0.0505	0.0325	0.0432	0.0133	0.0367	0.0132	0.0578	0.0203	0.0445	0.0131	0.0356	0.0165	0.0399	0.0222
Q	0.0382	0.0203	0.0299	0.0132	0.0271	0.0242	0.0359	0.0415	0.0401	0.0132	0.0392	0.0144	0.0495	0.0233	0.0345	0.0122	0.0452	0.0197	0.0414	0.0235
R	0.0471	0.0226	0.0413	0.0177	0.0538	0.0359	0.0322	0.0272	0.048	0.0136	0.0439	0.0147	0.0451	0.0156	0.0449	0.0151	0.0486	0.019	0.074	0.0435
S	0.0888	0.0353	0.0852	0.019	0.098	0.0408	0.1659	0.0677	0.0842	0.0153	0.0803	0.0193	0.1211	0.0194	0.0685	0.0153	0.075	0.0221	0.0679	0.0269
T	0.0579	0.0227	0.0606	0.0187	0.052	0.0293	0.0874	0.0545	0.055	0.0129	0.0539	0.0139	0.0643	0.0161	0.0582	0.0135	0.0494	0.016	0.0559	0.0269
V	0.0579	0.0209	0.0683	0.017	0.0599	0.0332	0.0578	0.0304	0.0522	0.0122	0.054	0.0143	0.0441	0.0129	0.0708	0.0144	0.0514	0.0174	0.0659	0.0304
W	0.0109	0.0093	0.0177	0.0097	0.0115	0.0149	0.0091	0.0108	0.011	0.0067	0.0125	0.0084	0.0074	0.0058	0.0109	0.0077	0.0082	0.0078	0.0064	0.0107
Y	0.0346	0.017	0.0408	0.015	0.0432	0.0317	0.0285	0.0227	0.0351	0.0114	0.037	0.0132	0.0303	0.0127	0.0339	0.0122	0.027	0.0135	0.0305	0.0199

Tabla A2-II .- Muestra los valores de frecuencia (Media y Desviación Estándar) para cada aminoácido en todos los conjuntos generados a partir del vector de características de dos dimensiones.

AA´s	Global		Set1		Set2		Set3		Set4		Set5		Set6		Set7		Set8		Set9	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	0.0562	0.026	0.0684	0.0175	0.068	0.0218	0.0482	0.0282	0.0446	0.0117	0.0473	0.0179	0.055	0.0231	0.0716	0.0383	0.0563	0.0259	0.0542	0.0483
C	0.016	0.0153	0.0139	0.0085	0.0183	0.0108	0.0328	0.0203	0.0141	0.009	0.0145	0.01	0.0116	0.0092	0.001	0.0043	0.0196	0.0178	0.0287	0.0338
D	0.0528	0.0225	0.0608	0.0141	0.0401	0.0141	0.0269	0.018	0.0629	0.0143	0.0624	0.0192	0.0527	0.0191	0.0567	0.0288	0.0512	0.0241	0.0224	0.022
E	0.0603	0.0275	0.0673	0.0145	0.0433	0.0147	0.029	0.0195	0.0671	0.0163	0.0796	0.0269	0.0498	0.0179	0.0764	0.0354	0.0624	0.0296	0.0278	0.0293
F	0.0482	0.0244	0.0433	0.0129	0.0601	0.0189	0.0817	0.0347	0.0448	0.0131	0.042	0.0156	0.0351	0.0126	0.0344	0.0188	0.0446	0.0246	0.0673	0.0455
G	0.0505	0.0236	0.0664	0.0159	0.066	0.0214	0.0397	0.0262	0.0443	0.0123	0.0405	0.016	0.0448	0.0195	0.0557	0.0287	0.0508	0.0241	0.0404	0.0423
H	0.0224	0.014	0.0223	0.009	0.0209	0.0114	0.0257	0.0206	0.0235	0.0087	0.0206	0.0097	0.0235	0.0117	0.0181	0.0135	0.023	0.015	0.0251	0.0342
I	0.0651	0.0219	0.064	0.014	0.0738	0.0185	0.072	0.0327	0.0685	0.0136	0.065	0.0178	0.0528	0.0154	0.0532	0.0217	0.0639	0.023	0.0685	0.0454
K	0.0724	0.0297	0.0734	0.0158	0.0519	0.0168	0.0494	0.0265	0.0772	0.0157	0.0886	0.0254	0.0603	0.0213	0.0951	0.0425	0.0787	0.0315	0.0573	0.0541
L	0.0971	0.0289	0.0837	0.0131	0.1043	0.019	0.1272	0.0376	0.0932	0.0133	0.1116	0.023	0.0745	0.0211	0.0855	0.0301	0.0911	0.026	0.1175	0.0554
M	0.0229	0.0125	0.0213	0.0083	0.0258	0.0101	0.0264	0.0148	0.0203	0.0069	0.0215	0.0095	0.0203	0.0091	0.0222	0.0141	0.0237	0.0129	0.0362	0.0298
N	0.0567	0.0232	0.0489	0.0127	0.0462	0.0141	0.0428	0.0205	0.0677	0.0151	0.059	0.0181	0.0774	0.0307	0.0539	0.0276	0.0572	0.0251	0.0427	0.0335
P	0.0438	0.0195	0.0442	0.0126	0.0416	0.0137	0.0434	0.0241	0.0442	0.0135	0.036	0.0137	0.057	0.0223	0.0403	0.0213	0.0447	0.0216	0.0505	0.0392
Q	0.0382	0.0203	0.0344	0.0115	0.0308	0.0119	0.0271	0.0205	0.0408	0.0135	0.0413	0.015	0.0485	0.0265	0.0455	0.0258	0.0419	0.0215	0.0288	0.0395
R	0.0471	0.0226	0.0456	0.0153	0.0409	0.0162	0.0513	0.0288	0.0463	0.0125	0.046	0.0165	0.0406	0.0185	0.0562	0.0297	0.0544	0.0257	0.0503	0.0517
S	0.0888	0.0353	0.069	0.0148	0.0821	0.018	0.1061	0.0465	0.0884	0.0151	0.0763	0.0191	0.1359	0.039	0.0799	0.0321	0.084	0.0302	0.118	0.0704
T	0.0579	0.0227	0.0574	0.0132	0.0612	0.0166	0.0515	0.0245	0.056	0.0127	0.0508	0.0132	0.0749	0.0361	0.0566	0.0242	0.0586	0.0218	0.0634	0.0467
V	0.0579	0.0209	0.0686	0.0146	0.0671	0.0152	0.0622	0.0289	0.0506	0.0114	0.052	0.0144	0.047	0.0157	0.0594	0.0279	0.0606	0.024	0.0568	0.0364
W	0.0109	0.0093	0.0121	0.0074	0.0177	0.0089	0.0175	0.0121	0.0108	0.0062	0.0114	0.0075	0.008	0.0056	0.0091	0.0106	0.001	0.0027	0.006	0.0146
Y	0.0346	0.017	0.0351	0.0121	0.0401	0.0136	0.0392	0.0248	0.0345	0.0109	0.0334	0.0124	0.0303	0.0131	0.0291	0.018	0.0324	0.018	0.0383	0.0378

Tabla A2-III .- Muestra los valores de frecuencia (Media y Desviación Estándar) para cada aminoácido en todos los conjuntos generados a partir del vector de características de tres dimensiones.

AA´s	Global		Set1		Set2		Set3		Set4		Set5		Set6		Set7		Set8		Set9	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	0.0562	0.026	0.073	0.0175	0.0456	0.0119	0.0536	0.0209	0.0563	0.0276	0.0542	0.0228	0.05	0.0174	0.0532	0.0516	0.0712	0.0401	0.052	0.0307
C	0.016	0.0153	0.015	0.0089	0.0141	0.009	0.0121	0.0093	0.0218	0.0183	0.013	0.0119	0.0169	0.0087	0.0293	0.0389	6E-05	0.0006	0.0337	0.0201
D	0.0528	0.0225	0.0553	0.0156	0.0639	0.0141	0.054	0.0175	0.0465	0.0245	0.0643	0.024	0.0496	0.0169	0.0206	0.0253	0.0553	0.0286	0.0266	0.018
E	0.0603	0.0275	0.0596	0.0158	0.07	0.0159	0.0515	0.017	0.057	0.031	0.0915	0.0298	0.0531	0.0179	0.0246	0.0319	0.0725	0.0352	0.0299	0.0208
F	0.0482	0.0244	0.0457	0.015	0.0451	0.0127	0.0354	0.0121	0.0486	0.03	0.0358	0.0155	0.0561	0.0182	0.0688	0.049	0.036	0.0217	0.0785	0.0349
G	0.0505	0.0236	0.0712	0.017	0.046	0.0125	0.0455	0.0194	0.049	0.0266	0.0451	0.0197	0.0462	0.0178	0.0404	0.047	0.0556	0.0284	0.041	0.0275
H	0.0224	0.014	0.0215	0.0093	0.0233	0.0085	0.0237	0.0113	0.024	0.0177	0.0197	0.0114	0.0218	0.0095	0.026	0.0385	0.0186	0.0138	0.0252	0.0205
I	0.0651	0.0219	0.0659	0.0151	0.0675	0.0132	0.0539	0.0145	0.0648	0.0271	0.0575	0.0179	0.0753	0.0178	0.0683	0.0467	0.0555	0.0239	0.0681	0.0324
K	0.0724	0.0297	0.0667	0.0178	0.0782	0.0155	0.0613	0.02	0.0747	0.0363	0.1024	0.0261	0.0638	0.0193	0.0569	0.0569	0.0888	0.043	0.0509	0.0284
L	0.0971	0.0289	0.0862	0.0136	0.0923	0.013	0.0763	0.02	0.0946	0.0318	0.0915	0.0249	0.1222	0.015	0.1205	0.0582	0.0857	0.0307	0.1203	0.0401
M	0.0229	0.0125	0.0223	0.0089	0.0207	0.007	0.0201	0.0088	0.0243	0.0144	0.0211	0.0103	0.0236	0.0096	0.0414	0.0324	0.0227	0.015	0.0258	0.0147
N	0.0567	0.0232	0.0472	0.013	0.0649	0.0148	0.0777	0.0289	0.0553	0.0276	0.0561	0.019	0.0556	0.0179	0.0435	0.0371	0.0543	0.0282	0.0426	0.0208
P	0.0438	0.0195	0.0438	0.0126	0.0436	0.0131	0.0563	0.0202	0.0458	0.024	0.0372	0.0165	0.0385	0.0126	0.0519	0.0432	0.0406	0.0226	0.0449	0.0248
Q	0.0382	0.0203	0.0335	0.0117	0.0396	0.0126	0.049	0.0268	0.0406	0.0274	0.0429	0.0173	0.036	0.0127	0.0258	0.037	0.0447	0.0258	0.0268	0.0207
R	0.0471	0.0226	0.0433	0.0152	0.0469	0.0125	0.0409	0.0163	0.0524	0.0283	0.052	0.0223	0.041	0.0134	0.0571	0.0573	0.053	0.029	0.0529	0.0297
S	0.0888	0.0353	0.0711	0.0157	0.0847	0.0155	0.1308	0.0348	0.0907	0.0416	0.0722	0.0216	0.0832	0.0182	0.1124	0.0704	0.0862	0.0421	0.1094	0.0484
T	0.0579	0.0227	0.059	0.0141	0.0551	0.0125	0.072	0.0324	0.06	0.0275	0.0503	0.0158	0.0557	0.014	0.0564	0.0419	0.0611	0.0305	0.0551	0.0298
V	0.0579	0.0209	0.0708	0.0141	0.0521	0.0117	0.0468	0.0142	0.0603	0.0268	0.0529	0.0183	0.0578	0.0142	0.0564	0.0381	0.0596	0.0274	0.062	0.0296
W	0.0109	0.0093	0.0131	0.0081	0.0113	0.0064	0.0082	0.0056	0.0002	0.0011	0.0104	0.0077	0.0146	0.0085	0.009	0.018	0.0084	0.0103	0.0172	0.0119
Y	0.0346	0.017	0.0358	0.012	0.0351	0.011	0.031	0.0128	0.033	0.0206	0.0301	0.0132	0.0388	0.013	0.0375	0.0405	0.0302	0.0204	0.0372	0.0244

Tabla A2-IV .- Muestra los valores de frecuencia (Media y Desviación Estándar) para cada aminoácido en todos los conjuntos generados a partir del vector de características de cuatro dimensiones.

AA´s	Global		Set1		Set2		Set3		Set4		Set5		Set6		Set7		Set8		Set9	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	0.0562	0.026	0.0495	0.0282	0.0466	0.0144	0.0639	0.0182	0.0683	0.0439	0.0553	0.0221	0.0567	0.0242	0.0527	0.0503	0.0554	0.0283	0.0698	0.0398
C	0.016	0.0153	0.0334	0.0185	0.015	0.0079	0.0149	0.0086	0.0226	0.0249	0.0114	0.0085	0.0135	0.0124	0.0258	0.032	0.0233	0.0242	0	0
D	0.0528	0.0225	0.0258	0.0172	0.0562	0.0157	0.0595	0.0153	0.0423	0.0271	0.0542	0.0179	0.0622	0.0237	0.0143	0.0221	0.0459	0.0245	0.0547	0.0285
E	0.0603	0.0275	0.0294	0.0202	0.0614	0.0193	0.0647	0.0159	0.0549	0.0395	0.0514	0.0188	0.0809	0.0295	0.0175	0.0308	0.056	0.031	0.0709	0.0343
F	0.0482	0.0244	0.0776	0.0345	0.0502	0.0159	0.0461	0.0139	0.0571	0.0356	0.0352	0.0129	0.0376	0.0164	0.0663	0.0497	0.0495	0.0314	0.0372	0.0239
G	0.0505	0.0236	0.0402	0.026	0.0439	0.0151	0.0638	0.0172	0.0509	0.0329	0.046	0.0188	0.0503	0.0227	0.0368	0.0501	0.0487	0.0269	0.0552	0.0285
H	0.0224	0.014	0.0298	0.0192	0.0222	0.0085	0.0223	0.0087	3E-05	0.0004	0.0231	0.0108	0.0213	0.0115	0.0331	0.0429	0.0256	0.0179	0.0206	0.0132
I	0.0651	0.0219	0.0701	0.0323	0.0721	0.015	0.0659	0.014	0.0593	0.0318	0.0526	0.0136	0.0596	0.0188	0.075	0.0494	0.0648	0.0271	0.0556	0.0243
K	0.0724	0.0297	0.0505	0.0266	0.0722	0.0196	0.0707	0.0174	0.0735	0.0471	0.0605	0.0205	0.0912	0.0291	0.0519	0.0532	0.0739	0.0369	0.088	0.0436
L	0.0971	0.0289	0.1197	0.0389	0.1102	0.0176	0.0862	0.0122	0.1086	0.0408	0.0746	0.0198	0.0908	0.0243	0.1208	0.0598	0.0955	0.0346	0.0861	0.0309
M	0.0229	0.0125	0.0261	0.0145	0.0215	0.0079	0.0219	0.0081	0.0291	0.0186	0.0201	0.0088	0.0216	0.011	0.0451	0.0349	0.0241	0.0142	0.0228	0.0153
N	0.0567	0.0232	0.0436	0.0214	0.0622	0.0172	0.0519	0.0141	0.0469	0.0279	0.0777	0.029	0.0551	0.0197	0.0431	0.0381	0.0554	0.0278	0.0544	0.0291
P	0.0438	0.0195	0.0453	0.0252	0.0401	0.0123	0.0442	0.0122	0.0441	0.0292	0.0566	0.0204	0.0404	0.0179	0.0482	0.0446	0.0462	0.0248	0.0412	0.0221
Q	0.0382	0.0203	0.0291	0.0212	0.0381	0.0117	0.0347	0.0118	0.0326	0.0311	0.0488	0.0264	0.0424	0.0192	0.0248	0.0339	0.0401	0.0275	0.0448	0.0262
R	0.0471	0.0226	0.0536	0.0285	0.0428	0.0117	0.0446	0.0139	0.0431	0.031	0.0396	0.0163	0.0526	0.0221	0.066	0.0637	0.0525	0.0292	0.0535	0.0286
S	0.0888	0.0353	0.1044	0.0446	0.087	0.0169	0.0734	0.0149	0.1061	0.0603	0.1321	0.0368	0.0742	0.0219	0.1128	0.072	0.0906	0.0407	0.0858	0.0417
T	0.0579	0.0227	0.0554	0.0274	0.0559	0.0125	0.057	0.0127	0.0545	0.0328	0.0747	0.0341	0.0519	0.0171	0.0607	0.0434	0.0597	0.0286	0.0605	0.0296
V	0.0579	0.0209	0.0603	0.0272	0.053	0.0127	0.0652	0.0146	0.0647	0.0337	0.0478	0.015	0.0569	0.0197	0.0546	0.0394	0.0595	0.0271	0.0595	0.0271
W	0.0109	0.0093	0.017	0.0116	0.0126	0.0075	0.0132	0.0077	0.012	0.0134	0.0079	0.0052	0.01	0.0076	0.0105	0.0195	5E-06	0.0001	0.0085	0.0105
Y	0.0346	0.017	0.0393	0.0246	0.0367	0.0116	0.0361	0.0116	0.0293	0.0268	0.0305	0.0124	0.0308	0.0134	0.0399	0.0398	0.0333	0.021	0.0309	0.0211

APÉNDICE 3

Tabla A3-I.- Descripción en función de términos de Proceso Biológico de Gene Ontology de cada una de las clases obtenidas mediante el vector de características de tres dimensiones.

	Set 1: 1116 proteínas (856 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0019538	Protein metabolism	22.31%		GO:0030435	Sporulation	1.99%
GO:0006082	Organic acid metabolism	17.17%		GO:0006364	RNA processing	1.75%
GO:0009308	Amine metabolism	13.08%		GO:0006897	Endocytosis	1.64%
GO:0006519	Aminoacid and derivative metabolism	12.50%		GO:0007165	Signal transduction	1.52%
GO:0006066	Alcohol metabolism	8.53%		GO:0007005	Mitochondrion organization and biogenesis	1.40%
GO:0006950	Response to stress	7.94%		GO:0045333	Cellular respiration	1.40%
GO:0006629	Lipid metabolism	6.78%		GO:0006113	Fermentation	1.40%
GO:0006766	Vitamin metabolism	5.37%		GO:0006800	Oxygen and reactive oxygen species metabolism	1.29%
GO:0006732	Coenzyme metabolism	5.37%		GO:0006312	Mitotic recombination	1.29%
GO:0009117	Nucleotide metabolism	4.67%		GO:0006081	Aldehide metabolism	1.29%
GO:0046483	Heterocycle metabolism	4.56%		GO:0006913	Nuclecytoplasmic transport	0.93%
GO:0006811	Ion transport	4.44%		GO:0042493	Response to drug	0.93%
GO:0006886	Intracellular protein transport	4.09%		GO:0006073	Glucan metabolism	0.93%
GO:0006092	Main pathways of carbohydrate metabolism	4.09%		GO:0016311	Dephosphorylation	0.93%
GO:0015849	Organic acid transport	3.62%		GO:0006730	One-carbon compound metabolism	0.93%
GO:0048193	Golgi vesicle transport	3.39%		GO:0006118	Electron transport	0.82%
GO:0006725	Aromatic compound metabolism	3.39%		GO:0030004	Monovalent inorganic cation homeostasis	0.82%
GO:0015837	Amine transport	3.39%		GO:0006401	RNA catabolism	0.70%
GO:0006790	Sulfur metabolism	2.80%		GO:0006352	Transcription initiation	0.70%
GO:0006399	tRNA metabolism	2.22%		GO:0006383	Transcription from RNA polymerase III promoter	0.70%
GO:0008643	Carbohydrate transport	2.22%		GO:0008219	Cell death	0.70%
GO:0006366	Transcription from RNA polymerase II promoter	1.99%				
	Set 2: 1446 proteínas (1075 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0006139	Nucleobase, nucleoside, nucleotide and nucleic acid metabolism	45.58%		GO:0007059	Chromosome segregation	2.05%
GO:0016043	Cell organization and biogenesis	32.09%		GO:0006944	Membrane fusion	1.49%
GO:0007049	Cell cycle	18.51%		GO:0009108	Coenzyme biosynthesis	1.21%
GO:0006464	Protein modification	15.44%		GO:0006887	Exocytosis	1.21%
GO:0050896	Response to stimulus	15.35%		GO:0008202	Steroid metabolism	0.93%
GO:0006793	Phosphorus metabolism	6.60%		GO:0046165	Alcohol biosynthesis	0.93%
GO:0007154	Cell comunication	5.58%		GO:0006752	Group transfer coenzyme metabolism	0.74%
GO:0000746	Conjugation	3.26%		GO:0006112	Energy reserve metabolism	0.65%
GO:0043037	Translation	2.88%		GO:0048278	Vesicle docking	0.56%
GO:0030437	Sporulation (sensu Fungi)	2.70%		GO:0006515	Misfolded or incompletely synthesized protein catabolism	0.47%
GO:0009101	Glycoprotein biosynthesis	2.42%				

Tabla A3-I (cont.)

	Set 3: 613 proteínas (443 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0007275	Development	20.32%		GO:0006913	Nucleocytoplasmic transport	2.93%
GO:0007028	Cytoplasm organization and biogenesis	17.38%		GO:0006073	Glucan metabolism	2.48%
GO:0050896	Response to stimulus	15.80%		GO:0006812	Cation transport	2.26%
GO:0007049	Cell cycle	15.12%		GO:0006402	mRNA catabolism	2.03%
GO:0019219	Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	13.77%		GO:0006644	Phospholipid metabolism	1.81%
GO:0007010	Cytoskeleton organization and biogenesis	12.19%		GO:0006112	Energy reserve metabolism	1.35%
GO:0007154	Cell communication	11.51%		GO:0006875	Metal ion homeostasis	1.35%
GO:0006313	DNA transposition	9.71%		GO:0008219	Cell death	1.35%
GO:0016192	Vesicle-mediated transport	7.67%		GO:0051049	Regulation of transport	1.13%
GO:0006997	Nuclear organization and biogenesis	7.22%		GO:0031123	RNA 3'-end processing	1.13%
GO:0007047	Cell wall organization and biogenesis	7.22%		GO:0048308	Organelle inheritance	1.13%
GO:0040007	Growth	7.00%		GO:0007109	Cytokinesis, completion of separation	0.68%
GO:0006796	Phosphate metabolism	5.87%		GO:0006379	mRNA cleavage	0.68%
GO:0000746	Conjugation	4.74%				
	Set 4: 593 proteínas (296 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0006412	Protein biosynthesis	16.22%		GO:0006875	Metal ion homeostasis	2.03%
GO:0006396	RNA processing	11.15%		GO:0051028	mRNA transport	2.03%
GO:0006366	Transcription from RNA polymerase II promoter	9.80%		GO:0006753	Nucleoside phosphate metabolism	2.03%
GO:0042254	Ribosome biogenesis and assembly	9.80%		GO:0006800	Oxygen and reactive oxygen species metabolism	1.69%
GO:0006950	Response to stress	9.12%		GO:0006944	Membrane fusion	1.35%
GO:0007126	Meiosis	4.39%		GO:0043094	Metabolic compound salvage	1.35%
GO:0009117	Nucleotide metabolism	4.05%		GO:0006289	Nucleotide-excision repair	1.35%
GO:0006811	Ion transport	3.72%		GO:0006100	Tricarboxylic acid cycle intermediate metabolism	1.35%
GO:0007047	Cell wall organization and biogenesis	3.04%		GO:0006360	Transcription from RNA polymerase I promoter	1.35%
GO:0045333	Cellular respiration	3.04%		GO:0006631	Fatty acid metabolism	1.35%
GO:0006511	Ubiquitin-dependent protein catabolism	3.04%		GO:0008535	Cytochrome c oxidase complex assembly	1.01%
GO:0007017	Microtubule-based process	2.70%		GO:0051052	Regulation of DNA metabolism	1.01%
GO:0006725	Aromatic compound metabolism	2.70%		GO:0006614	SRP-dependent cotranslational protein targeting to membrane	1.01%
GO:0006457	Protein folding	2.70%		GO:0007050	Cell cycle arrest	1.01%
GO:0005996	Monosaccharide metabolism	2.36%		GO:0030503	Regulation of cell redox homeostasis	1.01%
GO:0007020	Microtubule nucleation	2.36%		GO:0007121	Bipolar bud site selection	1.01%
GO:0045814	Negative regulation of gene expression, epigenetic	2.36%		GO:0006914	Autophagy	1.01%
GO:0006626	Protein targeting to mitochondrion	2.03%		GO:0000002	Mitochondrial genome maintenance	0.68%
GO:0030005	Di-, tri-valent inorganic cation homeostasis	2.03%				

Tabla A3-I (cont.)

	Set 5: 611 proteínas (426 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0019538	Protein metabolism	36.38%		GO:0006800	Oxygen and reactive oxygen species metabolism	2.58%
GO:0007028	Cytoplasm organization and biogenesis	23.00%		GO:0006261	DNA-dependent DNA replication	2.58%
GO:0016070	RNA metabolism	18.54%		GO:0007005	Mitochondrion organization and biogenesis	2.35%
GO:0006351	Transcription, DNA-dependent	12.44%		GO:0007114	Cell budding	2.11%
GO:0006605	Protein targeting	7.28%		GO:0050658	RNA transport	1.88%
GO:0000067	DNA replication and chromosome cycle	7.04%		GO:0030437	Sporulation (sensu Fungi)	1.88%
GO:0006325	Establishment and/or maintenance of chromatin architecture	6.10%		GO:0006732	Coenzyme metabolism	1.64%
GO:0016192	Vesicle-mediated transport	4.93%		GO:0007033	Vacuole organization and biogenesis	1.64%
GO:0006091	Generation of precursor metabolites and energy	4.93%		GO:0007127	Meiosis	1.41%
GO:0000087	M phase of mitotic cell cycle	3.52%		GO:0007264	Small GTPase mediated signal transduction	1.41%
GO:0006974	Response to DNA damage stimulus	3.05%		GO:0006109	Regulation of carbohydrate metabolism	0.94%
GO:0007059	Chromosome segregation	2.82%		GO:0000086	G2/M transition of mitotic cell cycle	0.94%
GO:0030468	Establishment of cell polarity (sensu Fungi)	2.58%				
	Set 6: 1018 proteínas (665 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0006810	Transport	31.79%		GO:0006281	DNA repair	1.97%
GO:0006996	Organelle organization and biogenesis	19.69%		GO:0016125	Sterol metabolism	1.69%
GO:0006464	Protein modification	13.64%		GO:0016044	Membrane organization and biogenesis	1.55%
GO:0006629	Lipid metabolism	11.81%		GO:0006914	Autophagy	1.41%
GO:0009628	Response to abiotic stimulus	6.05%		GO:0000070	Mitotic sister chromatid segregation	1.41%
GO:0000003	Reproduction	5.91%		GO:0007131	Meiotic recombination	1.13%
GO:0006355	Regulation of transcription, DNA-dependent	5.63%		GO:0043414	Biopolymer methylation	0.98%
GO:0006082	Organic acid metabolism	4.78%		GO:0000271	Polysaccharid biosynthesis	0.84%
GO:0030003	Cation homeostasis	4.22%		GO:0006270	DNA replication initiation	0.84%
GO:0007165	Signal transduction	3.80%		GO:0006515	Misfolded or incompletely synthesized protein catabolism	0.84%
GO:0009101	Glycoprotein biosynthesis	3.66%		GO:0006743	Ubiquinone metabolism	0.70%
GO:0006511	Ubiquitin-dependent protein catabolism	3.52%		GO:0007062	Syster chromatide cohesion	0.70%
GO:0000074	Regulation of progression through cell cycle	3.23%		GO:0006267	Pre-replicative complex formation and maintenance	0.70%
GO:0042158	Lipoprotein biosynthesis	3.23%		GO:0006308	DNA catabolism	0.56%
GO:0008380	RNA splicing	2.95%		GO:0007091	Mitotic metaphase/anaphase transition	0.56%
GO:0006399	tRNA metabolism	2.81%		GO:0000080	G1 phase of mitotic cell cycle	0.56%
GO:0006461	Protein complex assembly	2.39%		GO:0042168	Heme metabolism	0.56%
GO:0045333	Cellular respiration	2.25%		GO:0006298	Mismatch repair	0.56%
GO:0000910	Cytokinesis	2.25%		GO:0000288	mRNA catabolism, deadenylylation-dependent decay	0.56%
GO:0006944	Membrane fusion	1.97%		GO:0042401	Biogenic amine biosynthesis	0.56%

Tabla A3-I (cont.)

	Set 7: 222 proteínas (22 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0006412	Protein biosynthesis	30.43%		GO:0006812	Cation transport	8.70%
GO:0010038	Response to metal ion	13.04%		GO:0006091	Oxidative phosphorilation	4.35%
GO:0015986	ATP synthesis coupled proton transport	8.70%		GO:0006605	Protein targeting	4.35%
	Set 8: 545 proteínas (369 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0019538	Protein metabolism	40.11%		GO:0007059	Chromosome segregation	2.44%
GO:0042254	Ribosome biogenesis and assembly	8.67%		GO:0007033	Vacuole organization and biogenesis	2.44%
GO:0007010	Cytoskeleton organization and biogenesis	8.13%		GO:0007032	Endosome organization and biogenesis	2.17%
GO:0045045	Secretory pathway	5.69%		GO:0006818	Hydrogen transport	1.90%
GO:0009628	Response to abiotic stimulus	5.15%		GO:0006092	Main pathways of carbohydrate metabolism	1.63%
GO:0006357	Regulation of transcription from RNA polymerase II promoter	5.15%		GO:0030471	Spindle pole body and microtubule cycle (sensu Fungi)	1.63%
GO:0006323	DNA packaging	5.15%		GO:0006383	Transcription from RNA polymerase III promoter	1.63%
GO:0008380	RNA splicing	4.34%		GO:0007034	Vacuolar transport	1.63%
GO:0048193	Golgi vesicle transport	3.79%		GO:0000096	Sulfur amino acid metabolism	1.08%
GO:0006119	Oxidative phosphorilation	3.25%		GO:0030641	Hydrogen ion homeostasis	1.08%
GO:0006913	Nucleocytoplasmic transport	2.71%		GO:0007088	Regulation of mitosis	1.08%
GO:0017038	Protein import	2.71%				
	Set 9: 536 proteínas (38 sin Unknown)					
Término	**Descripción**	**%**		**Término**	**Descripción**	**%**
GO:0006412	Protein biosynthesis	28.95%		GO:0046467	Membrane lipid biosynthesis	5.26%
GO:0006626	Protein targeting to mitochondrion	7.89%		GO:0000028	Ribosomal small subunit assembly and maintenance	5.26%
GO:0000749	Response to pheromone during conjugation with cellular fusion	7.89%		GO:0006450	Regulation of translational fidelity	5.26%
GO:0016197	Endosome transport	7.89%		GO:0007105	Cytokinesis, site selection	2.63%
GO:0045333	Cellular respiration	5.26%				

APENDICE 4

PROGRAMAS UTILIZADOS

Software Comercial

Data Engine 2.10.012

Es una herramienta de análisis de datos elaborada por la empresa MIT (Management Intelligenter Techmologien) en Alemania. que funciona sobre plataforma Windows. Permite hacer diversos análisis y conversiones matemáticas y estadísticas con las tablas de datos e incluye métodos de análisis como perceptrón multicapa, redes de Kohonen, redes de Kohonen difusas, etc. Incluye herramientas para la visualización de los mapas de neuronas y la graficación de resultados.

Utilizamos esta herramienta para obtener las clasificaciones de nuestras secuencias a partir de las matrices (Ξ_1-Ξ_4), tanto utilizando el algoritmo Fuzzy c-means, como los Mapas Autoorganizados.

Prism 4.0 for Windows

Es un paquete de análisis bioestadístico, elaborado por Graph Pad Software, Inc. San Diego California. Incluye las herramientas de estadística descriptiva e inferencial más utilizadas en investigación científica y permite la graficación de los resultados.

Este programa se utilizó para casi todos los análisis estadísticos realizados en este trabajo, desde las pruebas de bondad de ajuste, las pruebas de normalidad y los análisis utilizando Mann Whitney, Kruskal-Wallis y Wilcoxon.

Software Desarrollado

Todas las herramientas de software necesarias para la generación de los vectores de características, detección de identidad entre clases, obtención de estadísticos simples, obtención del estimador Q para todas las clases y verificación de las clases, fueron programados utilizando Perl sobre plataforma Linux. Las herramientas desarrolladas fueron:

1) ***mapeo_lambda.pl***.- A partir de un archivo de secuencias de aminoácidos en formato FASTA (fig.5), genera las matrices Ξ_1-Ξ_4, haciendo los cálculos de frecuencias de aminoácidos, posición promedio y desviación promedio de la posición para la secuencia ordenada linealmente, bidimensionalmente, tridimensionalmente, etc. (utilizando el procedimiento descrito en la sección 3.2.2 de la metodología). Como resultado genera una matriz compuesta por tantos vectores de características como secuencias de entrada había en el archivo original.

2) ***find_identity.pl***.- Recibe como entrada dos archivos con una lista de identificadores de proteínas y nos arroja como resultado un nuevo archivo con el listado de identificadores presentes en ambos archivos. Este programa lo utilizamos en la sección 4.2.3 para determinar el porcentaje de identidad entre conjuntos generados a partir de los distintos métodos de clasificación.

3) ***fasta_subset.pl***.- Recibe como entrada un archivo con identificadores de proteínas y utiliza el archivo FASTA con todas las secuencias de proteínas de la especie, para generar un nuevo archivo en formato FASTA con las secuencias de las proteínas cuyos identificadores se incluyen en el archivo de entrada. Este programa lo utilizamos para poder regresar a los datos originales de secuencia de las proteínas y hacer los cálculos estadísticos de longitud y frecuencia para cada uno de los conjuntos obtenidos.

4) ***longitud.pl*** y ***frecuencia.pl***.- Ambos reciben como entrada un archivo en formato FASTA con la secuencia de un conjunto de proteínas y generan como salida un archivo con los valores de longitud de cadena, media y desviación estándar, para el primer caso, y un archivo con las frecuencias de cada aminoácido en cada proteína, además de la media y desviación estándar del conjunto. Estos programas se utilizaron para hacer comparaciones entre conjuntos en la sección 4.3.1 de los resultados.

5) ***valida.pl***.- Este programa es quizá el más complejo de los que elaboré para este trabajo. Para su operación se requiere tener los archivos de asociación de Gene Ontology para Saccharomyces (sc.bp) y la definición del DAG de GO para la Ontología (en nuestro caso process.ontology), para poder reconstruir el sub-DAG de GO para nuestra especie. Recibe como entrada un archivo con los identificadores de proteínas de alguno de nuestros conjuntos y su objetivo es obtener los valores del índice Q para cada uno de los términos significativos en nuestro conjunto de genes de entrada. Utiliza los siguientes módulos de perl: GO::OntologyProvider::OntologyParser, Graph::Directed, Set::Scalar, Math::Trig y Statistics::Descriptive.

Comienza por hacer una reconstrucción del DAG de GO para la ontología de proceso biológico, que incluya solamente los términos y relaciones correspondientes a nuestra especie. A continuación, ejecuta un algoritmo en el que va recorriendo el DAG iniciando por las hojas y para cada término de GO, si hay más de dos genes de nuestro conjunto de entrada que se encuentren presentes en la definición del término, hace los cálculos del índice Q (según se explicó en la sección 3.4.1 de la metodología). El siguiente paso consiste en eliminar todos los nodos terminales (que ya fueron analizados), pasando sus genes a los nodos padres. El proceso continúa del mismo modo hasta que sólo nos queda la raíz del DAG.

Como resultado nos genera un archivo que incluye el valor del índice Q para todos los términos de GO significativos en ese conjunto.

Este programa lo utilizamos para calcular los valores del índice Q que utilizamos en las comparaciones de la sección 4.3.2 de los resultados.

6) ***ajuste_definición.pl***.- Este programa se elaboró para poder determinar el ajuste de un conjunto de identificadores de genes a una definición dada como identificadores de términos de GO. Requiere los mismos módulos que el programa anterior y su mecánica es también semejante al inicio. Recibe como entrada una lista de identificadores de términos de GO, que corresponden a la definición que hemos elaborado para una clase dada y otro archivo con los identificadores de genes de uno de nuestros conjuntos que queremos contrastar con la definición.

Comienza nuevamente por la generación del sub-DAG de GO para nuestra especie y a continuación lee cada uno de los genes de nuestro conjunto de entrada y va recorriendo el DAG desde las hojas hacia arriba, determinando si el gen en

cuestión está incluído en alguna de las definiciones proporcionadas en el archivo de entrada, en algún nivel del DAG.

Como resultado genera un archivo con el porcentaje de genes de nuestro conjunto de entrada asociado a cada uno de los términos de la definición proporcionada, así como el porcentaje de genes no incluidos en la defínción.

Este programa lo utilizamos en la sección 4.4 de los resultados para verificar el ajuste de nuestros subconjuntos a las definiciones de cada una de las clases.

Páginas WEB utilizadas

Además de elaborar programas propios y utilizar programas comerciales, la obtención de datos y algunos análisis preliminares lo realizamos utilizando las herramientas y las bases de datos disponibles de manera gratuita en algunos sitios de Internet.

SGD Saccharomyces Genome Database

La mejor fuente de información genética y de biología molecular para *Saccharomyces cerevisiae* es el sitio Web de SGD (http://www.yeastgenome.org), mantenido por la Escuela de Medicina de la Universidad de Stanford. Es un sitio abierto, que proporciona toda su información (bases de datos y herramientas) de manera gratuita al público en general.

De este sitio fue de donde obtuvimos el archivo en formato FASTA con la secuencia de todos los genes de S. *cerevisiae*. También incluye algunas herramientas de análisis, como el GO Term Finder y el GO Slim Mapper que utilizamos para hacer una exploración preliminar sobre nuestros conjuntos.

Gene Ontology

El Proyecto Gene Ontology (GO) es un esfuerzo colaborativo que trata de contribuir a la necesidad de los biólogos de contar con descriptores consistentes de productos génicos entre las distintas bases de datos. El proyecto, que se enmarca dentro de los esfuerzos Open Source, que pretenden ofrecer información de manera abierta para su uso por investigadores en todo el mundo, inició como una colaboración entre las bases de datos de tres organismos modelo (FlyBase (Drosophila), Saccharomyces Genome Database (SGD) y Mouse Genome Database (MGD)) en 1998 y desde entonces ha crecido incluyendo mucha bases de datos de genomas de plantas, animales y microorganismos. En su sitio Web (http://geneontology.org) pueden obtenerse los archivos que describen las ontologías, así como archivos de asociación para todos los genomas incluidos en el proyecto. También contiene ligas a herramientas y utilerías relacionadas con GO.

Nosotros lo utilizamos para conocer los detalles de las ontologías y obtener los archivos de asociación necesarios para generar nuestro estimador Q.

AmiGO

La página http://www.godatabase.org proporciona una interfaz gráfica para la visualización de términos de GO o productos génicos en la ontología. Esta herramienta es particularmente útil cuando necesitamos visualizar la forma en que conjuntos de genes se disponen dentro de la ontología.

Utilizamos esta herramienta para identificar ramas del DAG que presentaban un alto grado de cobertura por los genes de nuestros conjuntos.

REFERENCIAS

1. Ackoff, R. L. (1996) On Learning and the Systems That Facilitate It. Center for Quality of Management Journal. 5 (2):27-35. http://cqmextra.cqm.org/cqmjournal.nsf/reprints/rp07300

2. Althschul SF, Carrol RJ, and Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology. 215 (410-

3. Ashburner, M, Ball, CA., Blake, JA., Botstein, D, Butler, H, Cherry, JM, Davis AP., Dolinski, K, Dwight, SS., Eppig, JT., Harris, MA., Hill, DP., Issel-Tarver, L, Kasarskis, A, Lewis, S, Matese, JC., Richardson, JE., Ringwald, M, Rubin, GM., and Sherlock, G (2000) Gene Ontology: tool for the unification of biology. Nat Genet. 25 (1):25-29.

4. Attwood, T. K., Beck, M. E., Bleasby, A. J., and Parry-Smith, D. J. (1-9-1994) PRINTS--a database of protein motif fingerprints. Nucl Acids Res. 22 (17):3590-3596.

5. Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., and Zygouri, C. (1-1-2003) PRINTS and its automatic supplement, prePRINTS. Nucl Acids Res. 31 (1):400-402.

6. Bairoch, A. (1991) PROSITE: a dictionary of sites and patterns in proteins. Nucl Acids Res. 19 (2241-2245.

7. Bairoch, A. (1993) The PROSITE dictionary of sites and patterns in proteins, its current status. Nucl Acids Res. 21 (13):3097-3103. http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=309737&blobtype=pdf

8. Bairoch, A. (1-9-1994) The ENZYME data bank. Nucl Acids Res. 22 (17):3626-3627.

9. Bateman, Alex, Birney, Ewan, Cerruti, Lorenzo, Durbin, Richard, Etwiller, Laurence, Eddy, Sean R., Griffiths-Jones, Sam, Howe, Kevin L., Marshall, Mhairi, and Sonnhammer, Erik L. L. (1-1-2002) The Pfam Protein Families Database. Nucl Acids Res. 30 (1):276-280.

10. Berry MJA and Linoff G (1996) Data Mining Techniques for Marketing, Sales and Customer Support. John Willey & Sons, Inc. USA.

11. Bezdek JC (1974) Cluster validity with fuzzy sets. Journal of Cybernetics. 3 (3):58-73.

12. Boisot, M. and Canals, A. (2004) Data, information and knowledge: have we got it right? Journal Of Evolutionary Economics. http://www.uoc.edu/in3/dt/20388/index.html

13. Bork P and Koonin E (1998) Predicting Functions from Protein Sequences. Where are the bottlenecks? Nature Genetics. 18 (313-318. http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v18/n4/abs/ng0498-313.html

14. Boyle, EI, Weng, S, Gollub, J, Jin, H, Botstein, D, Cherry, JM, and Sherlock, G (12-12-2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 20 (18):3710-3715.

15. Brenner S (2002) Life sentences: Ontology recapitulates philology. Genome Biology. 3 (4):comment1006.1-comment1006.2. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=11983049

16. Coello G (2006) MAGO: MicroArray Gene Ontology.

17. Corpet F (25-11-1988) Multiple sequence alignment with hierarchical clustering. Nucl Acids Res. 16 (22):10881-10890.

18. Corpet F, Gouzy J, and Kahn D (1998) The ProDom database of protein domain families. Nucl Acids Res. 26 (1):323-326. http://intl-nar.oxfordjournals.org/cgi/reprint/26/1/323

19. Corpet F, Servant F, Gouzy J, and Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucl Acids Res. 28 (1):267-269. http://intl-nar.oxfordjournals.org/cgi/reprint/28/1/267

20. Darden, Lindley, Tabery, and James (2005) Molecular Biology. Stanford Encyclopedia of Phylosophy. http://plato.stanford.edu/archives/spr2005/entries/molecular_biology

21. Enright AJ, Kunin V, and Ouzounis CA (2003) Protein families and TRIBES in genome sequence space. Nucl Acids Res. 31 (15):4632-4638. http://nar.oupjournals.org/cgi/reprint/31/15/4632

22. Feng, D. F. and Doolittle, R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal Of Molecular Evolution. 25 (4):351-360.

23. Ferran E, Pflugfelder B, and Ferrara P (1994) Self-organized neural maps of human protein sequences. Protein Science. 3 (507-521.

24. Gene Ontology Consortium (1-1-2004) The Gene Ontology (GO) database and informatics resource. Nucl Acids Res. 32 (90001):D258-D261.

25. Gerstein, M (2000) Integrative database analysis in structural genomics. Nature Structural Biology. 7 (960-963. http://www.nature.com/nsmb/journal/v7/n11s/abs/nsb1100_960.html

26. Goodenough, U (2006) Reductionism and Holism, Chance and Selection, Mechanism and Mind. Zygon. 40 (2):369-380.

27. Gracy J and Argos P (1-3-1998a) Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. Bioinformatics. 14 (2):164-173.

28. Gracy J and Argos P (1-3-1998b) Automated protein sequence database classification. II. Delineation Of domain boundaries from sequence similarities. Bioinformatics. 14 (2):174-187.

29. Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences of the United States of America. 84 (13):4355-4358.

30. Griffiths, A., Miller, J., Suzuki, D., Lewontin, R., and Gelbart, W. (2000) An Introduction to genetic analysis. W.H.Freeman and Co. New York. 860 pp. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=iga.TOC

31. Guralnik V and Karypis G (2001) A Scalable Algorithm for Clustering Protein Sequences. Workshop on Data Mining in Bioinformatics, BIOKDD. 73-80. http://www.cs.rpi.edu/~zaki/BIOKDD01/guralnik.ps.gz

32. Halkidi M, Batistakis Y, and Vazirgiannis M (2001) On clustering validation techniques. Journal of Intelligent Information Systems. 17 (2/3):107-145.

33. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, and Chen YZ (2004) Predicting Functional Family of Novel Enzymes irrespective of Sequence Similarity: A Statistical Learning Approach. Nucl Acids Res. 32 (21):6437-6444. http://intl-nar.oupjournals.org/cgi/content/full/32/21/6437

34. Henikoff, S. and Henikoff, J. G. (11-12-1991) Automated assembly of protein blocks for database searching. Nucl Acids Res. 19 (23):6565-6572.

35. Henikoff, S., Henikoff, J. G., and Pietrokovski, S. (1-6-1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics. 15 (6):471-479.

36. Higgins DG and Sharp PM (15-12-1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 73 (1):237-244.

37. Higgins DG, Thompson JD, and Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods in Enzymology. 266 (383-402.

38. Huang JY and Brutlag DL (2001) The EMOTIF database. Nucl Acids Res. 29 (1):202-204. http://nar.oxfordjournals.org/cgi/content/full/29/1/202?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=1&author1=Huang&andorexacttitle=and&andorexacttitleabs=and&andorexactfulltext=and&searchid=1118161420319_3400&stored_search=&FIRSTINDEX=0&sortspec=relevance&volume=29&firstpage=202&fdate=1/1/2001&tdate=12/31/2001&journalcode=nar

39. Jain AK, Murty MN, and Flynn PJ (1999) Data Clustering: A review. AMC Computing Surveys. 31 (3):264-323.

40. Johnson, M. S. and Doolittle, R. F. (1986) A method for the simultaneous alignment of three or more amino acid sequences. Journal Of Molecular Evolution. 23 (3):267-278.

41. Kanehisa M and Bork P (2003) Bioinformatics in the Post-sequence era. Nature Genetics (Suppl). 33 (305-310.

42. Kanehisa, M, Goto, S, Kawashima, S, and Nakaya, A (1-1-2002) The KEGG databases at GenomeNet. Nucl Acids Res. 30 (1):42-46.

43. Kohonen T (1990) The Self Organizing Map. Proceedings of the IEEE. 78 (9):1464-1480. http://ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=2115&arnumber=58325&count=7&index=2

44. Kuri-Morales AF, Galavíz-Casas J, and Herrera-Alcántara O (2006) Pattern-Based Data Compression and the MDL principle. Information and Computation. in press (

45. Kuri-Morales AF and Ortíz-Posadas MR (1-1-2006) A New Approach for Representation in Biological Sequences. WSEAS Transactions on Biology and Biomedicine. 3 (1):31-36.

46. Lehninger A (1982) Bioquímica. Ediciones Omega, S.A. Barcelona, España. 73.

47. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, and Bork P (1-1-2002) Recent improvements to the SMART domain-based sequence annotation resource. Nucl Acids Res. 30 (1):242-244.

48. Lewis SE (2004) Gene Ontology: looking backwards and forwards. Genome Biology. 6 (1):103-http://genomebiology.com/2004/6/1/103

49. Lipman DJ and Pearson WR (22-3-1985b) Rapid and sensitive protein similarity searches. Science. 227 (4693):1435-1441.

50. Lipman DJ and Pearson WR (1985a) Rapid and sensitive protein similarity searches. Science. 227 (1435-1441.

51. Mayr E (2006) Biology in the Twenty-First Century. BioScience. 50 (10):895-897.

52. Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. (1-1-1999) MIPS: a database for genomes and protein sequences. Nucl Acids Res. 27 (1):44-48.

53. MIT GmbH (1997) Data Engine. 2.10.012):

54. Needleman SB and Wunsh CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. Journal of Molecular Biology. 48 (443-453.

55. Nevill-Manning CG, Wu TD, and Brutlag DL (1998) Highly specific protein sequence motifs for genome analysis. Proceedings of the National Academy of Sciences of the United States of America. 95 (11):5865-5871. http://intl.pnas.org/cgi/reprint/95/11/5865

56. Ouzounis CA, Coulson R, Enright A, Kunin V, and Pereira-Leal J (2003) Classification schemes for protein structure and function. Nature Reviews Genetics. 4 (7):508-519.

57. Overbeek, R, Larsen, N, Pusch, GD., D'Souza, M, Evgeni S, Yrpides, N, Fonstein, M, Maltsev, N, and Selkov, E (1-1-2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucl Acids Res. 28 (1):123-125.

58. O´Neil, D. (2005) Classification of living things. Physical Anthropolgy tutorials. http://anthro.palomar.edu/animal/default.htm

59. Pasquier C, Pronponas VJ, and Hamodrakas SJ (2001) PRED-CLASS: Cascading Neural Networks for generalized protein classification and genome-wide applications. Proteins: Structure, Function and Genetics. 44 (361-369.

60. Rezaee RM., Lelieveldt BPF, and Reiber JHC (1998) A new cluster validity index for the fuzzy c-mean. Pattern Recognition Letters. 19 (3-4):237-246.

61. Schultz J, Milpetz F, Bork P, and Ponting CP (26-5-1998) SMART, a simple modular architecture research tool: Identification of signaling domains. PNAS. 95 (11):5857-5864.

62. Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, and Bucher P (2002) PROSITE: A documented database using patterns and profiles as motif descriptors. Briefings in Bioinformatics. 3 (3):265-274.

63. Smith TF and Waterman MS (1981) Identification of common molecular subsequences. Journal of Molecular Biology. 147 (195-197.

64. Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (15-9-2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucl Acids Res. 28 (18):3442-3444.

65. Sonnhammer ELL, Eddy SR, and Durbin R (2005) Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments. Proteins: Structure, Function and Genetics. 28 (3):405-420. http://www3.interscience.wiley.com/cgi-bin/fulltext/52461/PDFSTART

66. Tatusov, RL., Koonin, EV., and Lipman, DJ. (24-10-1997) A Genomic Perspective on Protein Families. Science. 278 (5338):631-637.

67. Taylor WR (19-12-1989) A flexible method to align large numbers of biological sequences. Journal Of Molecular Evolution. 28 (1-2):161-169.

68. Van Regenmortel, MHV (2004) Reductionism and complexity in molecular biology. EMBO Reports. 5 (11):1016-1020.

69. Whisstock JC and Lesk AM (2003) Prediction of Protein Function from Protein Sequence and Structure. Quarterly Reviews of Biophysics. 36 (3):307-340. http://journals.cambridge.org/bin/bladerunner?REQUNIQ=1113000045&REQSESS=4412868&118000REQEVENT=&REQINT1=197161&REQAUTH=0

70. Windham MP (1981) Cluster validity for Fuzzy Clustering algorithms. Fuzzy Sets and Systems. 5 (177-185.

71. Wu CH, Huang H, Nikolskaya A, Hu Z, Yeh LS, and Barker WC (2004) The iProClass Integrated database for protein functional analysis. Computational Biology and Chemistry. 28 (87-96.

72. Xie XL and Beni G (1991) A Validity Meassure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Inteligence. 13 (8):841-847.

73. Yona G (1999) Methods for Global Organization of the Protein Sequence Space. Ph.D., The Hebrew University, Jerusalem, Israel. http://citeseer.ist.psu.edu/78371.html

74. Yona G, Linial N, and Linial M (1999) Protomap: Automatic Classification of Protein Sequences A Hierarchy of Protein Famillies, and Local Maps of the Protein Space. Proteins: Structure, Function and Genetics. 37 (3):360-378. http://www3.interscience.wiley.com/cgi-bin/abstract/68502070/ABSTRACT

75. Zadeh LA (1965) Fuzzy Sets. Information Control. 8 (338-353.