Análisis de datos de Microarreglos (usos y espectatívas)

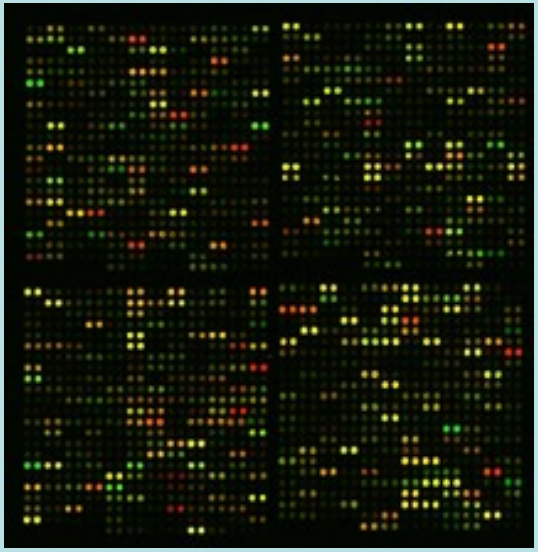
Gerardo Coello Coutiño
Unidad de Cómputo,
Instituto de Fisiología Celular,
UNAM

MICROARREGLOS DE DNA

Es una colección de oligos de DNA arreglados en una matriz o superficie sólida.

Cada spot se considera un gen individual.

Existen 2 tipos de microarreglos (canal único y doble canal)



Puede albergar hasta 40000 genes.

Generalmente son usados para medir el nivel de expresión de mRNA en miles de genes simultáneamente.

ANTECEDENTES

A más una década de su surgimiento, los microarreglos de DNA no han resultado ser la panacea que se creyó originalmente.

Los microarreglos de DNA pueden ser, -todavía-, una herramienta útil. El análisis simultáneo de miles de genes en condiciones experimentales puede ayudar en la prevención, diagnóstico y terapia de enfermedades.

Han permitido entender mejor enfermedades complejas, sobre todo aquellas que provocan cambios metabólicos generalizados.

COMPLEJIDAD FUNCIONAL

Un sistema de análisis de datos de microarreglos, debe incorporar información adicional, de tal forma que nos permita visualizar la complejidad de la condición experimental, enfatizando las interacciones y regulación que ocurren en los sistemas biológicos.

Considerar la expresión génica (mRNA) exclusivamente, **generalmente** nos lleva a conclusiones producto de hallazgos estadísticos.

ANALISIS DE DATOS DE MICROARREGLOS

No existe un estándar, generalmente se siguen un conjunto de pasos:

1) Preproceso filtrado de genes.
   a) Normalización de los datos.
   b) Eliminación de genes contradictorios, ruido, etc.

2) Obtención de un conjunto de genes sub/sobreexpresados.
   a) Gran cantidad de métodos, desde estadística simple hasta simulaciones Monte Carlo, pasando por redes neuronales.
   b) **genArise** (http://www.ifc.unam.mx/genarise)

3) Agrupamiento o “clustering” de acuerdo al nivel de expresión.
   a) Basados en distancias (k-means, c-means, vecinos cercanos, …) redes neuronales, mapas autoorganizados, FFT, etc.

4) Exploración funcional, no utilizan nivel de expresión.
   a) Utilizan clasificaciones funcionales como **Gene Ontology**.

FUENTES IMPORTANTES

Vías metabólicas y algo mas:

Reactome (Cold Spring Harbor) y **Pathways** (Kegg), son, acaso, las bases de datos mas importantes en rutas metabólicas, ligandos, enzimas, función biológica y otros chismes.

Vocabularios controlados y otras clasificaciones funcionales:

Gene Ontology (GO) . Los principales genomas han adoptado GO para anotar todos sus genes. Por otro lado, Uniprot hace anotaciones con GO a todos las proteínas curadas de su base de datos (GOA).

GENE ONTOLOGY (GO)

Base de conocimiento, fundamentadas en la evidencia, ya sea experimental, bibliográfica o inferida electrónicamente.

GO es una clasificación funcional basada en evidencias.
GO es jerárquico.
GO es una estructura artificial y su estructura misma puede ser modelada
GO es una gráfica dirigida acíclica (DAG)
GO es dinámica, se actualiza cada 30 minutos

GO es la clasificación funcional mas usada y mejor anotada.

GO describe 3 atributos (ontologías) de los productos génicos:

Molecular function (**MF**): Qué funciones tiene a nivel molecular?

Biological process (**BP**): A qué procesos biológicos contribuyen esas funciones?

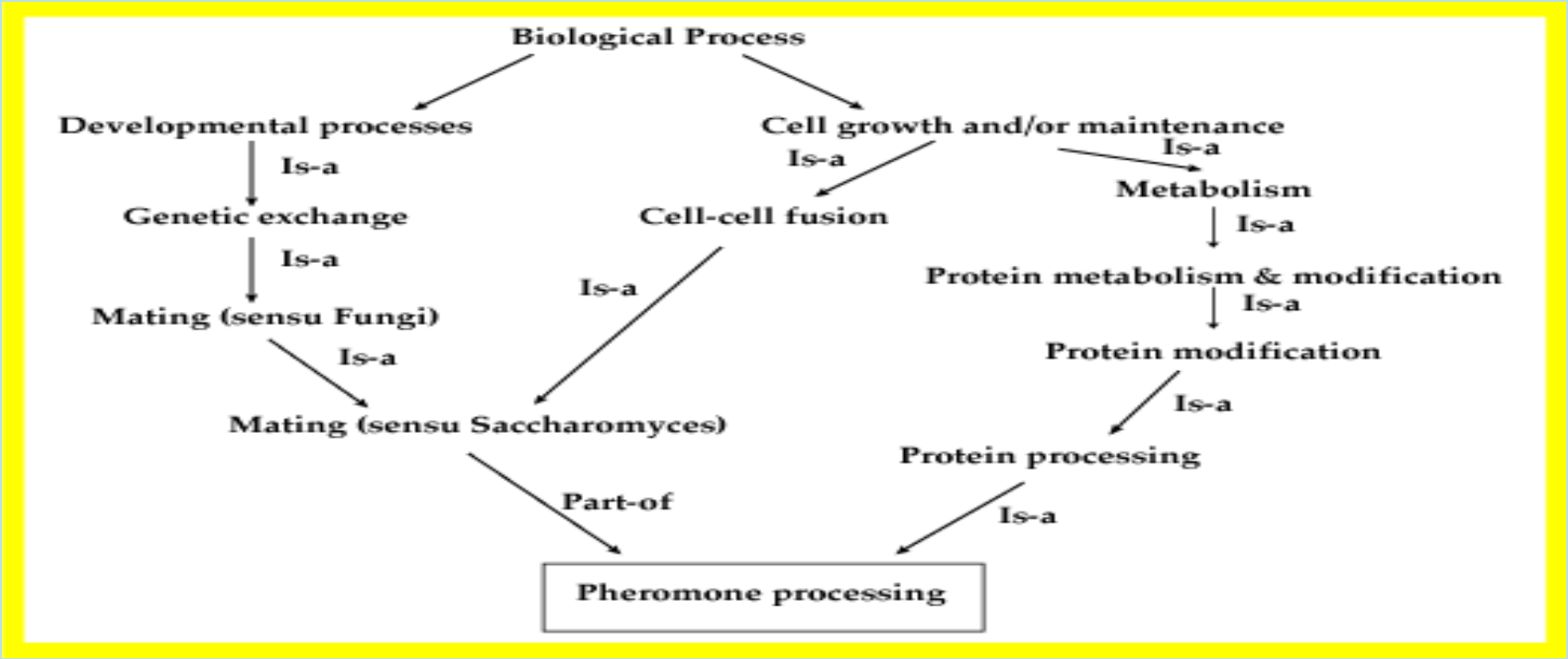
Celullar component (**CC**): Dónde esta localizado en la célula?

Los productos génicos participan también en otras instancias biológicas, complejas en si mismas:

a) Rutas metabólicas (pathways)
b) Regulación génica.
c) Interacciones proteína proteína, proteina-DNA, etc

1) Consiste de una red (DAG) cuyos **nodos** (términos) incluyen definiciones y descripciones funcionales de los genes.
2) Relaciones entre los **nodos**, principalmente:
   a) **is-a**
   b) **part-of**

3) Genes asociados a cada **nodo**, vía archivos de asociación, disponibles en los principales genomas eucariontes secuenciados



PROPUESTA PARA UN ANALISIS FUNCIONAL CON GENE ONTOLOGY (MAGO)

1) La población corresponde al genoma/microarreglo bajo estudio

2) La muestra de genes de entrada son todos aquellos genes del microarreglo que están **up**/**down** regulados.

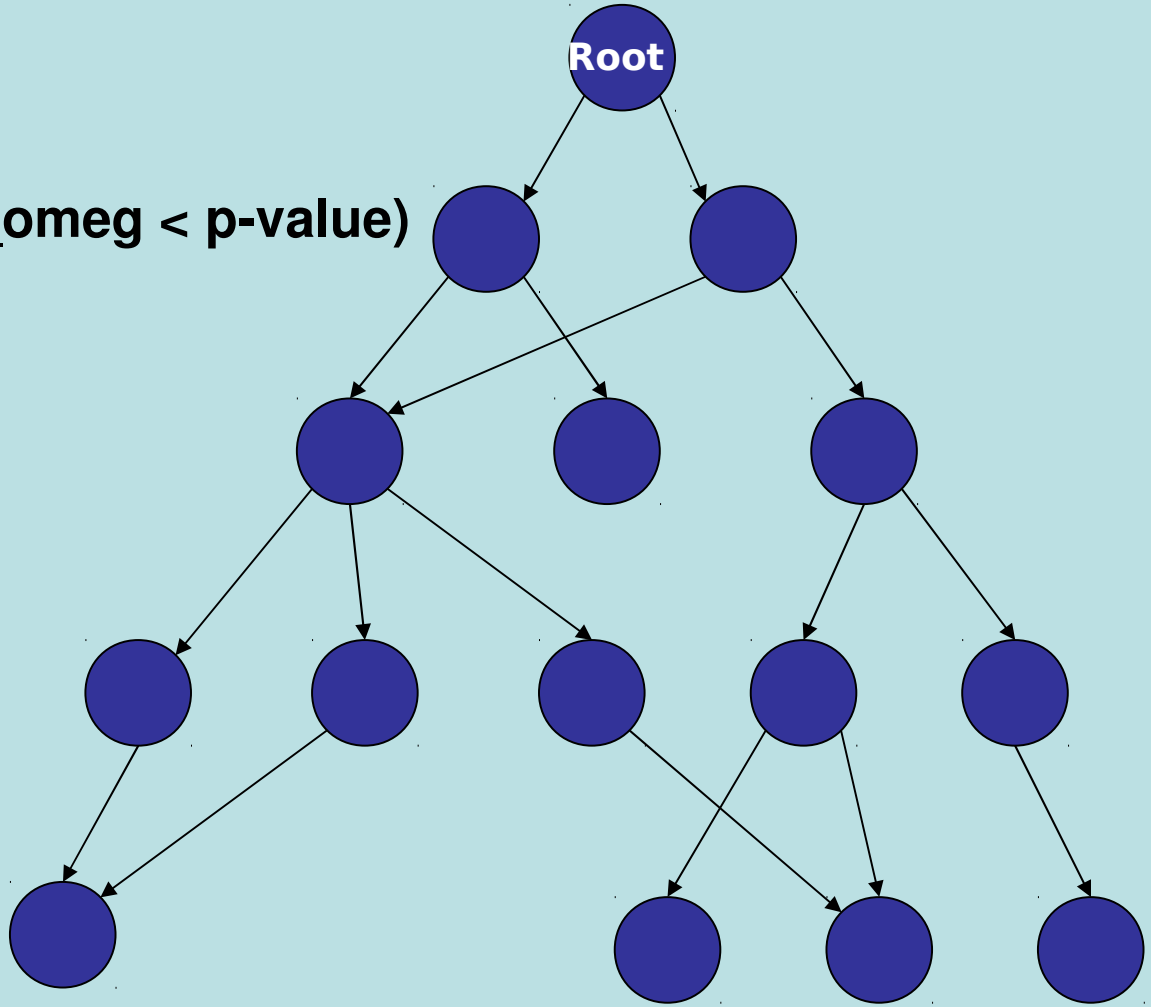
3) Para cada nodo de GO y para cada ontología (BP, MF, CC) evalúa:

a) Cuál es la representatividad de la muestra de entrada en el nodo?
Calcula indice de representatividad **omega.**

b) Cuál es la representatividad del nodo en la muestra de entrada?
Calcula indice de representatividad **rho.**

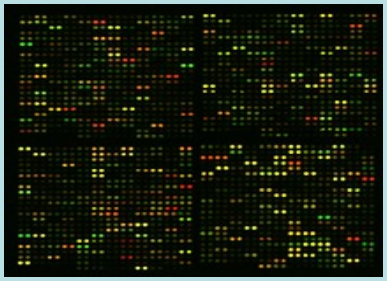
4) Para todos aquellos nodos sub/sobre representados **simultáneamente** bajo las condiciones **a** y **b**, evalúa una métrica (función de evaluación) , que depende de las variables **omega** y **rho** .

```
while (root) {
   hojas <- get_hojas
   foreach hoja (hojas) {
      calcula omega(hoja)
      calcula rho(hoja)
      if (pvalue_rho && pvalue_omeg < p-value)
         calcula f(omega , rho)
      }
   }
   remueve hoja
}
```



METODOLOGIA GENERAL

1) A partir de los datos de un microarreglo, obtener el conjunto de genes sobre/sub (UP/DOWN) expresados usando ***genArise*** (Z-score >= 1.5).



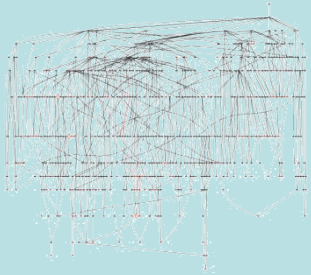
Genes sobre-expresados

Genes sub-expresados

2) Con los conjuntos UP/DOWN se hace una clasificación funcional usando Gene Ontology (MAGO p<=0.05) obtenemos los nodos significativos para BP y MF.

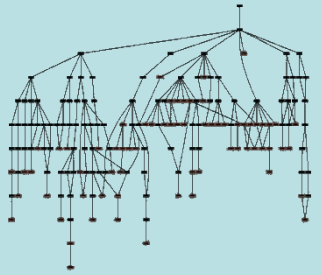
Genes sobre-expresados

Genes sub-expresados



BP

+

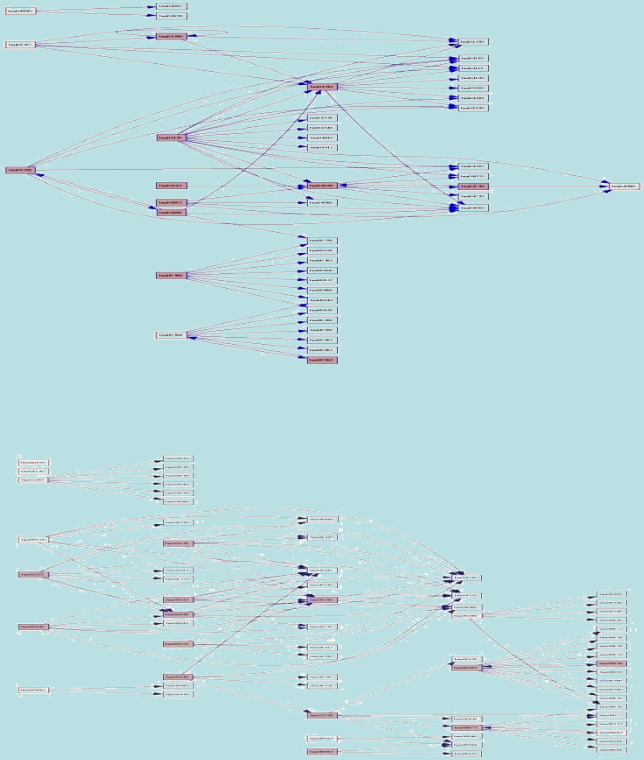


MF

3) Para cada conjunto UP/DOWN, la unión de genes de los nodos significativos BP + MF es el conjunto de entrada para obtener los pathways mejor representados (significativos) en KEGG, utilizando una variación del algoritmo usado en Gene Ontology.

{ Genes GO BP } + { Genes GO MF } ⟹ { Genes KEGG Pathways }

{ Genes KEGG Pathways } → UP / DOWN



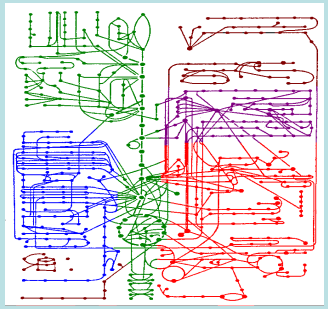
Pathways significativos

4) Para cada gráfica (UP/DOWN) de pathways significativos, obtenemos el pathway mas visitado aplicando el siguiente algoritmo:

```
APSP Floyd-Warshall
foreach shortest_path (u,v) {
        @vertices = path_vertices(shortest_path);
         foreach vertex (@vertices) {
           $mas_visitado{vertex}++
         }
}
mas_visitado = max( keys %mas_visitado)
```

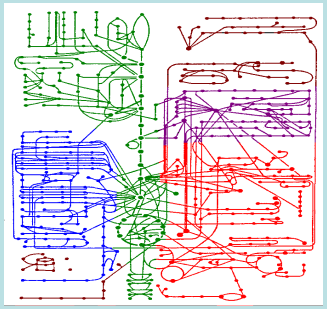
5) Cada pathway mas visitado genera un MetaPathway, que incluye a todos los pathways que inciden en él.

Pathway mas visitado

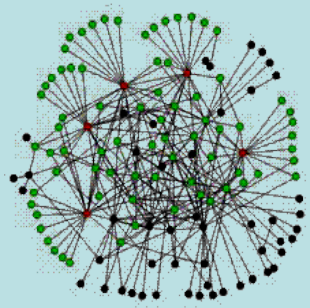


Gráfica MetaPathways

6) Genera una gráfica (metagenes) de todos genes presentes en el Pathway.



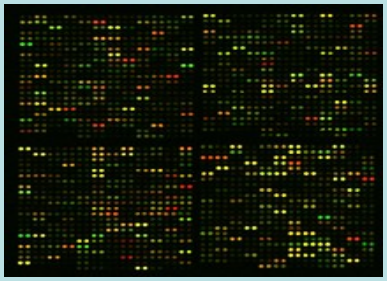
Gráfica MetaPathways



Red de genes

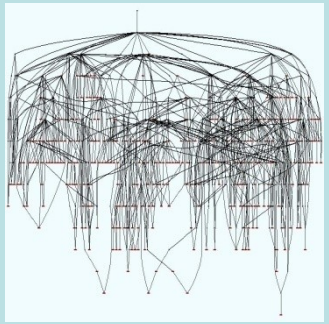
7) Se obtienen los genes **más visitados** de la red de genes

8) Se obtiene la gráfica final exclusiva de los genes **mas visitados** y todas las relaciones que existen entre ellos.

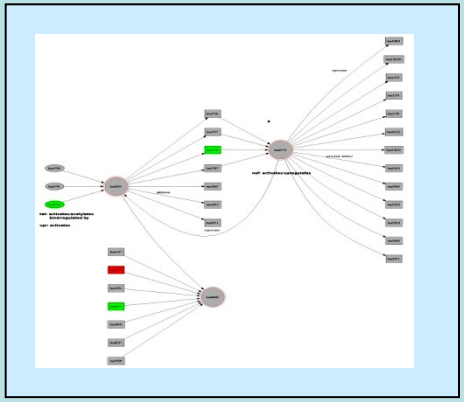


Microarreglo

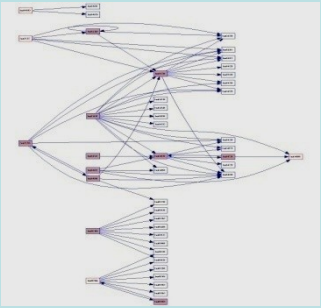
Análisis funcional con Gene Ontology



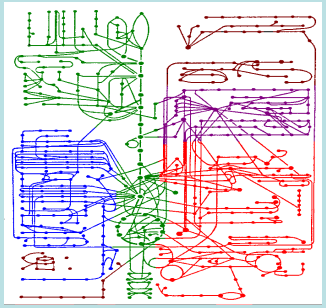
Pathways involucrados
(Ruta mas visitada)



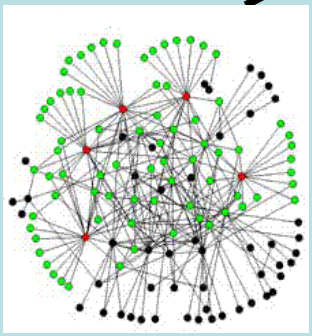
Genes relevantes
(mas visitados)



MetaPathways



Red de genes



VALIDACIÓN DEL MODELO

1)Obtenemos todos los genes de las vías metabólicas de diabetes (tipo I, tipo II, juvenil) y los genes de **todas** las rutas metabólicas que se relacionan de manera directa con estos pathways. Obtenemos APSP con estos genes y obtenemos una muestra con los más visitados y los menos visitados.

2)Estos genes los consideramos como entrada en MAGO y los clasificamos funcionalmente.

a) Reportados por MAGO MF (no BP)

b) Pahways encontrados con la muestra adhoc 300

c) Pathways a partir de una muestra más visitados 30-30

RESULTADOS

1) Partiendo de microarreglos de células hepáticas; 5 de individuos delgados y sanos y 4 individuos obesos y diabéticos. Se normalizaron y analizaron exclusivamente los 300 genes relacionados con diabetes.

a) No se encontró ningún gen específico a ninguna condición experimental

2) Se hizo el mismo análisis con todos los 23000 genes por cada microarreglo.

a) Se encontraron 5 no genes expresados en los individuos sanos y 32 genes no expresados en los individuos diabéticos (10 genes no siquiera anotados).

b) Se utilizaron estos genes como entrada de MAGO y se observó esto

3) Con los 27 genes encendidos/apagados se construyó una gráfica metagenes 27

GRACIAS POR SU ATENCION

Especialmente:

Dr. José Galaviz	Fac. Ciencias, UNAM
Antonio Coello Pérez	Fac. Ciencias, UNAM
MC. Ana María Escalante	IFC, UNAM
Dr. Jorge Ramírez	IFC, UNAM
Sergio Rojas	IFC,UNAM

PROPUESTA PARA UN ANALISIS FUNCIONAL CON GENE ONTOLOGY (MAGO)

1) La población corresponde al genoma bajo estudio

2) La muestra de genes de entrada son todos aquellos genes del microarreglo que están up/down regulados.

3) Para cada nodo de GO y para cada ontología (BP, MF, CC) evalúa:

a) Cuál es la representatividad de la muestra de entrada en el nodo?
Calcula indice de representatividad **omega.**

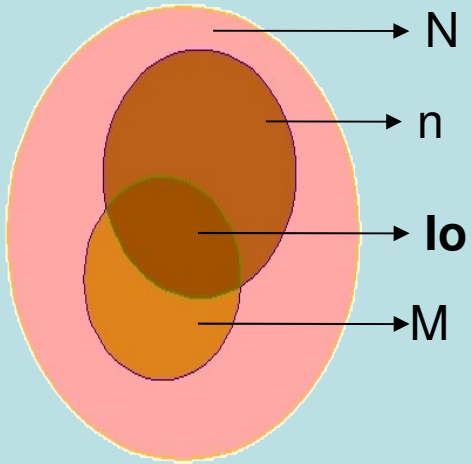
b) Cuál es la representatividad del nodo en la muestra de entrada?
Calcula indice de representatividad **rho.**

4) Para todos aquellos nodos sobrerepresentados **simultáneamente** bajo las condiciones **a** y **b**, evalúa una métrica (función de evaluación) , que depende de las variables **omega** y **rho** .

CALCULO DE OMEGA

$$\omega = \frac{Io - Ie_{(n)}}{M - Ie_{(n)}} \quad ;(Io - Ie_{(n)}) \geq 0,\ \ pvalue \leq 0.01$$

st: 143.723



Io = Número de genes de la muestra en el nodo

$$Ie_{(n)} = \frac{nIo}{N}$$

Número de genes que esperamos en un muestreo al azar de tamaño n

$$pvalue = P(i \geq Io) = \sum_{i=Io}^{M} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

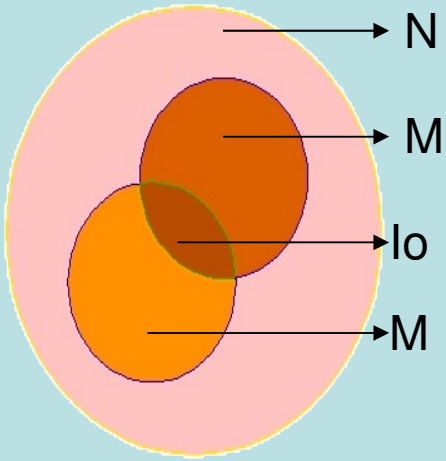
N = Número de genes en el genoma o microarreglo
n = Tamaño de la muestra
M = Número de genes asociados al nodo de GO

pvalue expectativa de encontrar cuando menos (**Io**) genes al azar

CALCULO DE RHO

$$\rho = \frac{Io - Ie_{(M)}}{M - Ie_{(M)}} \quad ;(Io - Ie_{(M)}) \geq 0,\ \ p - value \leq 0.01$$

ost 40.983



Io = Número de genes de la muestra en el nodo

$$Ie_{(M)} = \frac{M \cdot Io}{N}$$

Número de genes que esperamos en un muestreo al azar de tamaño M

$$pvalue = P(i \geq Io) = \sum_{i=Io}^{M} \frac{\binom{M}{i}\binom{N-M}{M-i}}{\binom{N}{M}}$$

N = Número de genes en el genoma o microarreglo
M = Tamaño de la muestra
M = Número de genes asociados al nodo de GO

pvalue expectativa de encontrar cuando menos (**Io**) genes al azar

Definimos **omega** y **rho** como componentes de un vector en el círculo unitario. Para cada nodo (n_i) de GO:

$$\mathbf{v}_{ni} = \left(\omega_{ni}, \rho_{ni}\right)$$

y en coordenadas polares:

$$\mathbf{v}_i = \left(\|\mathbf{v}_i\|, \theta\right) \qquad \text{donde :}$$

$$\|\mathbf{v}_i\| = \sqrt{\omega_{ni}{}^2 + \rho_{ni}{}^2} \text{ norma Euclidiana de } \mathbf{v}_i$$

$$\theta = \arctan \frac{\rho_{ni}}{\omega_{ni}}$$

Tanto mejores valores de un nodo cualquiera, cuanto la magnitud del vector **Vi** es máxima. Esto es, cuando los componentes de **Vi** se maximizan **simultáneamente.**

$$\omega_{ni} = \rho_{ni} = 1 \;\; ; \;\; \|\mathbf{v}_i\| = \sqrt{2} \;\; ; \;\; \theta = \frac{\pi}{4} = \frac{\rho_{ni}}{\omega_{ni}}$$

Sea $\gamma = \tan(2\theta - \pi/2) \in [-\pi/2, \pi/2]$

con la propiedad un cero en el valor óptimo $\theta = \pi/4$

La cual mapea a θ en todo $\Re$ cuando $[\theta = 0, \theta = \pi/2]$

Función de Evaluación

Necesitamos una función que nos entregue calificaciones exponencialmente mayores en aquellos nodos con mayor magnitud y equilibrio.

$$\|\mathbf{v}_i\| \cong \sqrt{2} \quad \text{y} \quad \theta \cong \pi/4$$

Se propone la siguiente función de evaluación:

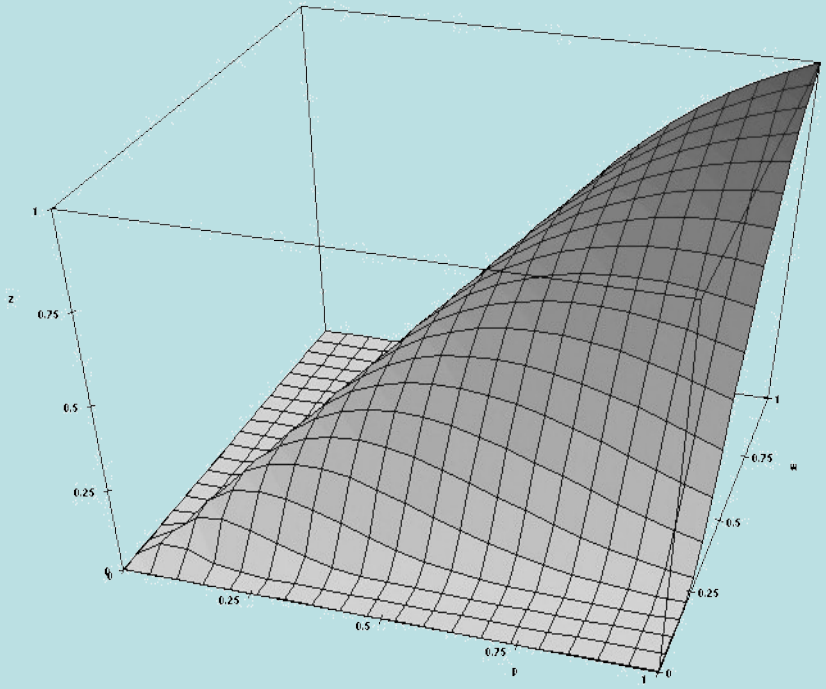
$$\mathbf{f}_i(\|\mathbf{v}_i\|, \theta) = \|\mathbf{v}_i\| e^{-\tan^2(\gamma_i)}$$

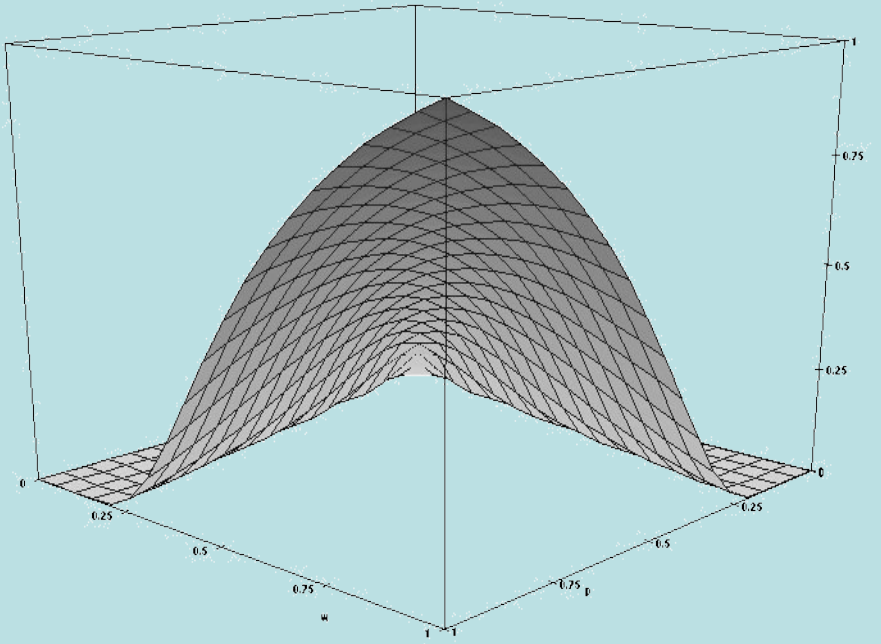
Efectuando manipulaciones trigonométricas y algebraicas tenemos que

$$\tan^2(\gamma_i) = \frac{\omega_{ni}^4 + \rho_{ni}^4}{4\omega_{ni}^2\rho_{ni}^2} - \frac{1}{2}$$

y sustituyendo en nuestra función de evaluación obtenemos finalmente

$$f_i(\rho_{ni}, \omega_{ni}) = \sqrt{\rho_{ni}^2 + \omega_{ni}^2}\, e^{\left(\frac{1}{2} - \frac{\omega_{ni}^4 + \rho_{ni}^4}{4\omega_{ni}^2\rho_{ni}^2}\right)}$$





```
while (root) {
   hojas <- get_hojas
   foreach hoja (hojas) {
      calcula omega(hoja)
      calcula rho(hoja)
      if (pvalue_rho && pvalue_omeg < p-value)
         calcula f(omega , rho)
      }
   }
   remueve hoja
}
```