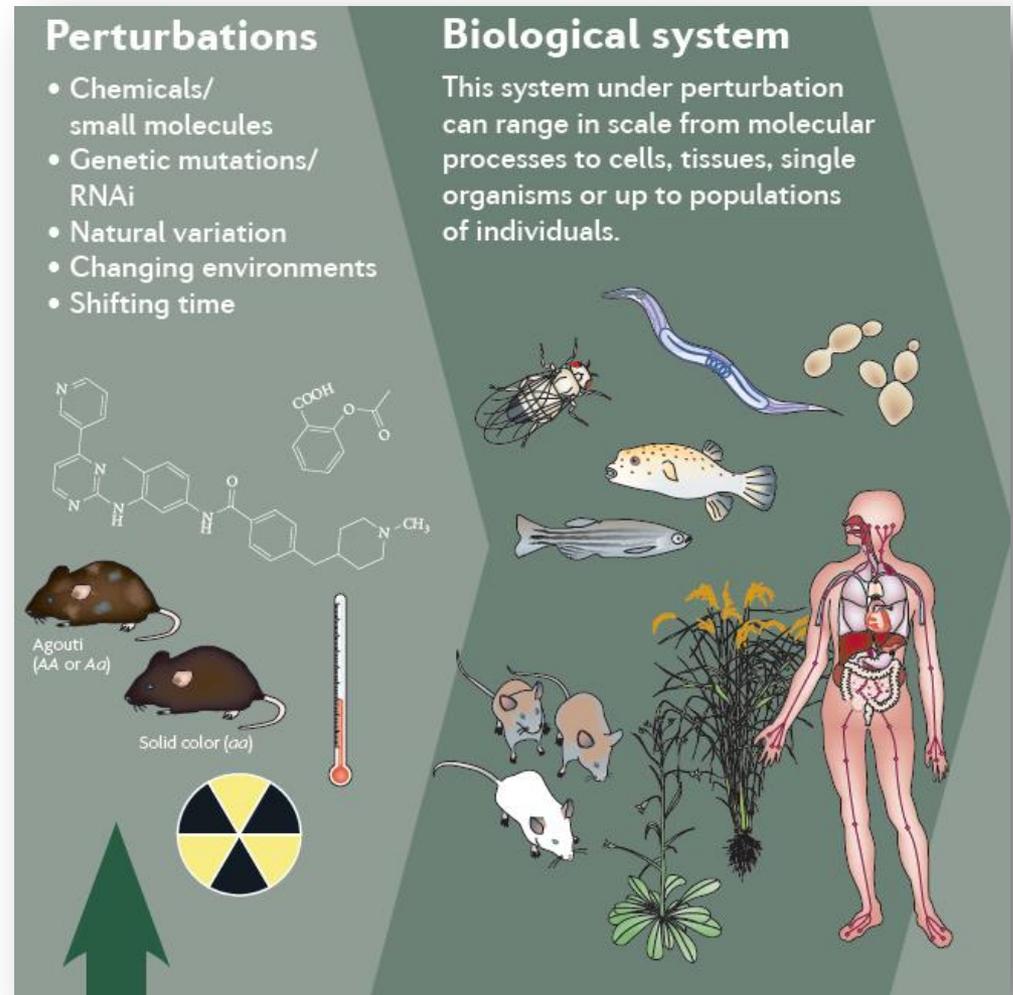


Redes. Clusters. Biologia.

Contexto: Biología de Sistemas Integrativa

Estudio de sistemas biológicos mediante

- Análisis de respuestas globales ante perturbaciones sistemáticas
- Integración de datos en modelos predictivos



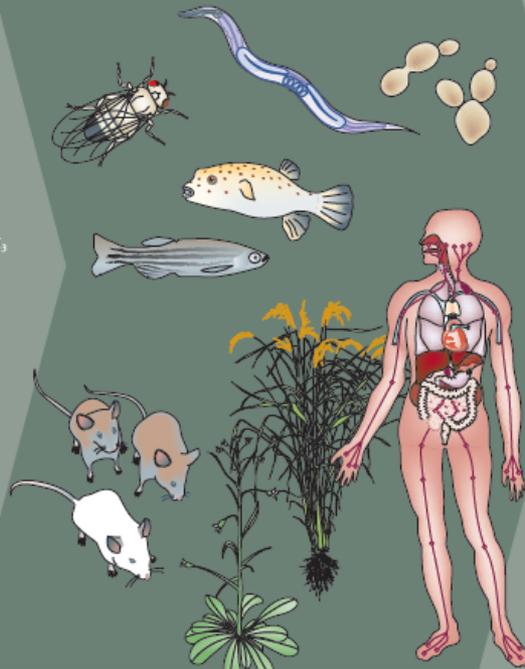
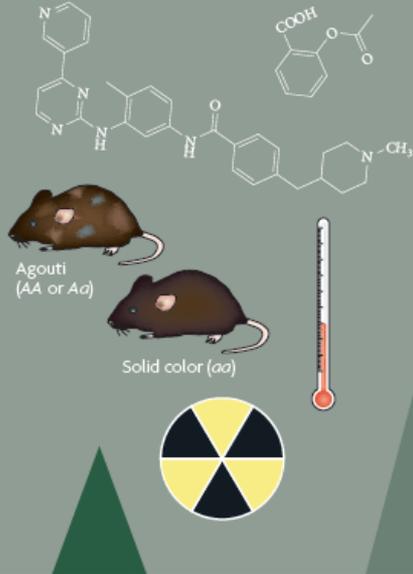
Biología de Sistemas Integrativa

Perturbations

- Chemicals/
small molecules
- Genetic mutations/
RNAi
- Natural variation
- Changing environments
- Shifting time

Biological system

This system under perturbation can range in scale from molecular processes to cells, tissues, single organisms or up to populations of individuals.



Molecular state measurements (nodes)

	Large-scale dataset types	Technologies
Genome	Whole-genome DNA sequences, SNPs and CNVs	DNA sequencing, genotyping microarrays
Epigenome	Chromatin modifications and structure	ChIP-seq, methyl-seq, DHS-chip
Transcriptome	Transcript abundances, translation rate and microRNAs	DNA microarrays, RNA-seq, CAGE, GRO-seq, ribosome profiling
Proteome	Protein abundances and modifications	NMR, mass spectrometry, multiparameter FACS
Metabolome	Metabolite profiling	Mass spectrometry, liquid chromatography

Molecular interaction measurements (edges)

	Data types	Technologies
Physical	Protein-protein	Immuno-precipitation (IP), co-affinity purification, yeast two-hybrid, protein arrays, kinase-substrate measurements
	Protein-DNA, protein-RNA	Genome-wide chromatin immuno-precipitation (ChIP), DNA binding arrays
	Protein-small molecule, reaction fluxes	Isotope labeling, mass spectrometry
Genetic and functional	Synthetic lethality, epistasis	Synthetic genetic arrays (SGA) combinatorial RNAi, population genetics
	Cause-effect relationships	Genetic perturbation (gene knockout, RNAi) followed by phenotyping (microarrays, cellular imaging); <i>trans</i> eQTLs (expression quantitative trait loci)

Biología de Sistemas Integrativa

Molecular state measurements (nodes)

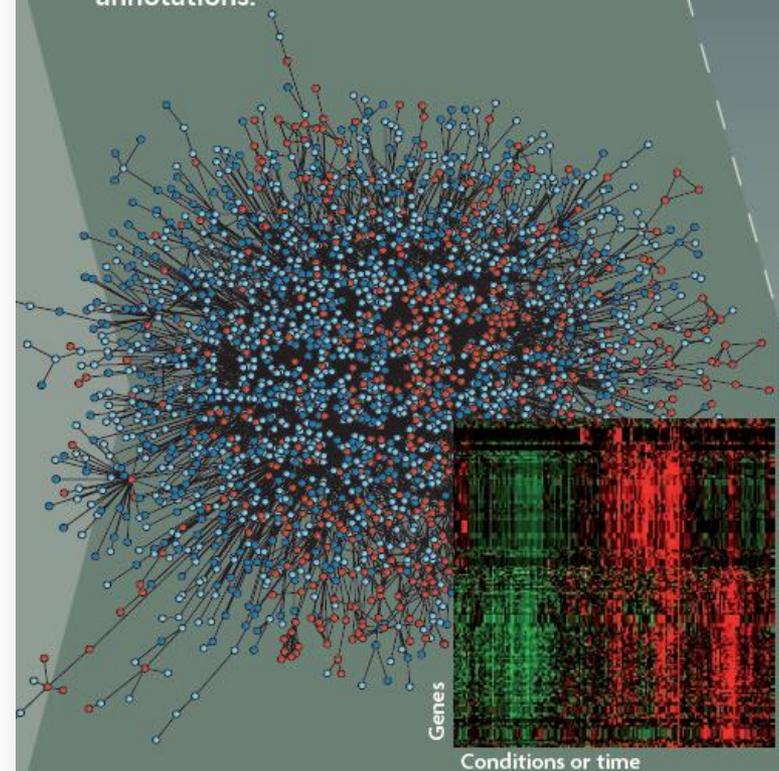
	Large-scale dataset types	Technologies
Genome	Whole-genome DNA sequences, SNPs and CNVs	DNA sequencing, genotyping microarrays
Epigenome	Chromatin modifications and structure	ChIP-seq, methyl-seq, DHS-chip
Transcriptome	Transcript abundances, translation rate and microRNAs	DNA microarrays, RNA-seq, CAGE, GRO-seq, ribosome profiling
Proteome	Protein abundances and modifications	NMR, mass spectrometry, multiparameter FACS
Metabolome	Metabolite profiling	Mass spectrometry, liquid chromatography

Molecular interaction measurements (edges)

	Data types	Technologies
Physical	Protein–protein	Immuno-precipitation (IP), co-affinity purification, yeast two-hybrid, protein arrays, kinase-substrate measurements
	Protein–DNA, protein–RNA	Genome-wide chromatin immuno-precipitation (ChIP), DNA binding arrays
	Protein–small molecule, reaction fluxes	Isotope labeling, mass spectrometry
Genetic and functional	Synthetic lethality, epistasis	Synthetic genetic arrays (SGA) combinatorial RNAi, population genetics
	Cause–effect relationships	Genetic perturbation (gene knockout, RNAi) followed by phenotyping (microarrays, cellular imaging); <i>trans</i> eQTLs (expression quantitative trait loci)

Molecular databases

New measurements are stored alongside existing data, including functional annotations.



	Annotations	States	Interactions
Selected molecular databases	GO , KEGG , REACTOME , NetPath , UCSC Genome Browser	Genbank , GEO , ArrayExpress , Proteome Commons	BioGRID , HPRD , IntAct , TRANSFAC , STRING , iHOP , functionalnet.org

Conocimiento Biológico

	Interaction type(s)	Detection methodologies	Databases
Physical	Protein-protein	Yeast-2-hybrid (Y2H) ¹⁻³ , co-immuno-precipitation (Co-IP) ⁴ , mass spectroscopy ^{5,6} , affinity purification coupled with mass spectroscopy (AP/MS) ^{7,8}	BioGRID ⁹ , IntAct ¹⁰ , APID ¹¹ , STRING ¹² , MINT ¹³ , DIP ¹⁴ , HPRD ¹⁵ , MIPS-MPPI ¹⁶ , Netpath ¹⁷ , DroiD ¹⁸
	Protein-DNA (e.g., regulatory networks)	Yeast-1-hybrid (Y1H) ¹⁹ , chromatin immuno-precipitation based methods (CHIP-CHIP) ²⁰ , DNA-footprinting ²¹	TRANSFAC ²² , UniProbe ²³ , DroiD ¹⁸ , BioGRID ⁹ , TcoF-DB ²⁴ , BIPA ²⁵ , hPDI ²⁶ , EDGEdb ²⁷ , NPIDB ^{28,29}
	Protein-RNA	RNA electro-mobility shift (RNA-EMSA) ³⁰ , RNA-pull down ³¹	PRID ³² , BIPA ²⁵ NPInter ³³ , RBPDB ³⁴ , PRIDB ³⁵ , StarBase ³⁶
	Metabolic (e.g., enzyme-substrate, ligand-receptor)	Mass spectroscopy based selective reaction monitoring (SRM) ^{6,37} , NMR ³⁸ , affinity purification ⁸ , co-IP ³ , fluorescence spectroscopy ³⁹	Reactome ⁴⁰ , KEGG ⁴¹ , BioCyc and MetaCyc ⁴² , HMDB ^{43,44} , EcoCyc ⁴⁵ , HumanCyc ⁴⁶ , ConsensusPathDB ⁴⁷
	Protein/gene-compound (e.g., drug-target, chemical-protein)	Chemical structure ^{48,49} , forward or reverse chemo-genomic/proteomic profiling ⁵⁰⁻⁵² , in silico predictions ⁵³	SuperTarget ⁵⁴ , Matador ⁵⁴ , DrugBank ⁵⁵ , ChemProt ⁵⁶ , STITCH ⁵⁷ , AffinDB ⁵⁸ , MatrixDB ⁵⁹ , PSMDB ⁶⁰ , PDB-Ligand ⁶¹ , ChEMBL ⁶² , ConsensusPathDB ⁴⁷
Functional	Genetic (gene-gene)	Synthetic genetic array (SGA) ⁶³ , Epistatic Miniarray Profiling (E-MAP) ⁶⁴ , co-expression profiling ^{65,66}	BioGRID ⁹ , DRYGIN ⁶⁷ , CYGD ⁶⁸ , DroiD ¹⁸ , ConsensusPathDB ⁴⁷
	Gene-Disease	Literature curation, clinical and sequence information	OMIM ⁶⁹ , HuDiNe ⁷⁰ , Disasome ⁷¹
Omics data type		Detection methodologies	Databases
Transcriptomics		Microarray ⁷² , RNASeq ⁷³	GEO ⁷⁴ , SMD ⁷⁵ , TCGA ⁷⁶ , GXD ⁷⁷ , ONCOMINE ⁷⁸ , ArrayExpress ⁷⁹
RNAi (phenomics)		RNAi interference assay ⁸⁰	RNAiDB ⁸¹ , GenomeRNAi ⁸² , siRecords ⁸³
Epigenomics		Methylation profiling ⁸⁴	DAnCER ⁸⁵ , DiseaseMeth ⁸⁶ , PubMeth ⁸⁷ , MethDB ⁸⁸ , MethCancerDB ⁸⁹ , MethyCancer ⁹⁰
Mutation / SNP		SNP Array ⁹¹ , genome sequencing ⁹²	TCGA ⁷⁶ , dbSNP ⁹³ , dbQSNP ⁹⁴ , GWAS Central ⁹⁵ , OMIM ⁶⁹
Proteomics		CHIP ^{11,2,96} , Mass Spectrometry ^{5,6}	PDB ⁹⁷ , ExPASy ⁹⁸ , InterPro ⁹⁹ , World-2DPAGE ¹⁰⁰ , JASPAR ¹⁰¹
Phosphorylation profile		Mass Spectrometry ⁵ , literature curation	PhosphoGRID ¹⁰² , PhosphoELM ¹⁰³ , PHOSIDA ¹⁰⁴

Portal de Conocimiento Biológico

The screenshot shows the Pathguide website interface. At the top, there is a navigation bar with the Pathguide logo and the tagline "the pathway resource list". Below the logo, there are navigation links for "Home", "BioPAX", and "cBio". A search bar is located in the top right corner. On the left side, there is a vertical navigation menu with categories such as "Navigation", "Protein-Protein Interactions", "Metabolic Pathways", "Signaling Pathways", "Pathway Diagrams", "Transcription Factors / Gene Regulatory Networks", "Protein-Compound Interactions", "Genetic Interaction Networks", "Protein Sequence Focused", "Other", "Search", "Organisms", "Availability", "Standards", "Analysis", "Statistics", "Database Interactions", and "Contact".

The main content area is titled "Complete Listing of All Pathguide Resources". It contains the following text:

Pathguide contains information about **547** biological pathway related resources and molecular interaction related resources. Click on a link to go to the resource home page or 'Details' for a description page. Databases that are free and those supporting BioPAX, CellML, PSI-MI or SBML standards are respectively indicated.

If you know of a pathway resource that is not listed here, or have other questions or comments, please [send us an e-mail](#).

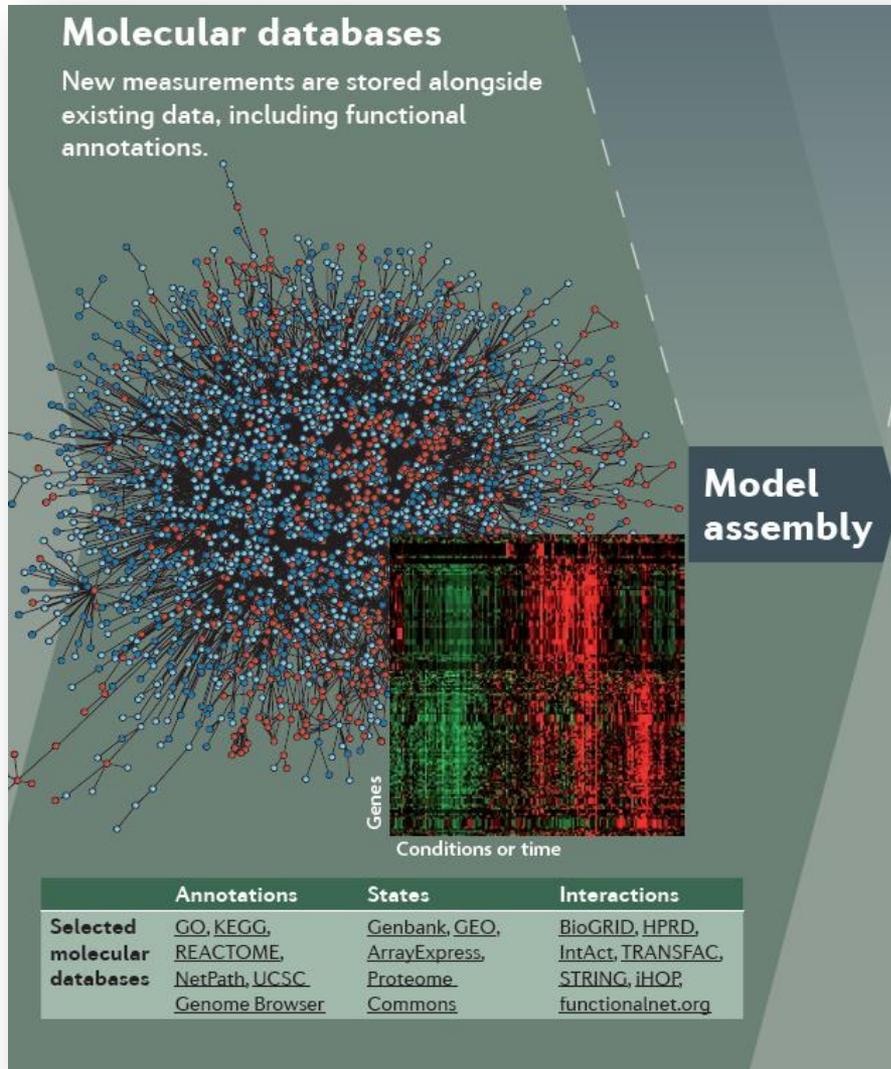
On the right side, there is a "News" section with two entries:

- Major new update of Pathguide** August 2010
We now have information about ~550 resources!
- Visual navigation added** May 2010
Click the 'Database Interactions' link on the

Below the news section, there is a section titled "Protein-Protein Interactions" with a sub-header "Database Name (Order: alphabetically | by web popularity)". This section contains a table of resources:

Database Name (Order: alphabetically by web popularity)	Full Record	Availability
2P2Idb - The Protein-Protein Interaction Inhibition Database	Details	Free
3D-Interologs - 3D-Interologs	Details	Free
3DID - 3D interacting domains	Details	Free
ADAN - Prediction of protein-protein interaction of modular domains	Details	X
AHD2.0 - Arabidopsis Hormone Database 2.0	Details	Free
AllFuse - Functional Associations of Proteins in Complete Genomes	Details	X
aMAZE - Protein Function and Biochemical Pathways Project	Details	X
ANAP - Arabidopsis Network Analysis Pipeline	Details	Free
AnimalTFDB - Animal Transcription Factor Database	Details	Free
AntiJen - AntiJen a Kinetic, Thermodynamic and Cellular Database	Details	Free
APID - Agile Protein Interaction DataAnalyzer	Details	Free
AS-ALPS - Alternative Splicing - induced ALteration of Protein Structure	Details	Free
ASD - Allosteric Database	Details	Free
ASEdb - Alanine Scanning Energetics Database	Details	Free
ASPD - Artificial Selected Proteins/Peptides Database	Details	Free
ATDB - Animal Toxin Database	Details	Free
AtPID - Arabidopsis thaliana Protein Interactome Database	Details	Free

Biología de Sistemas Integrativa



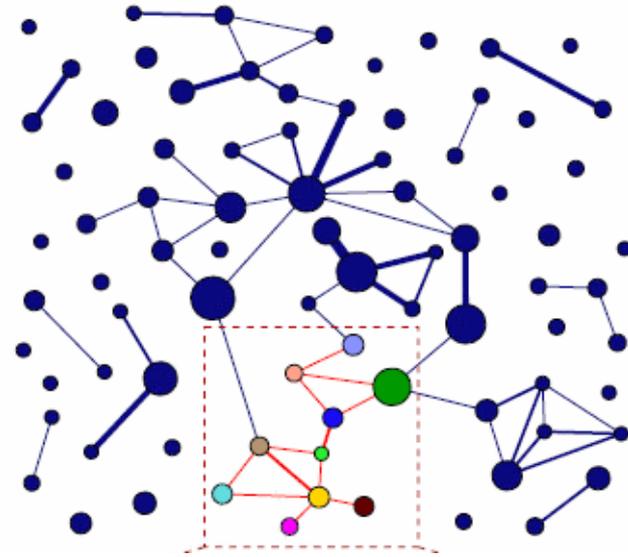
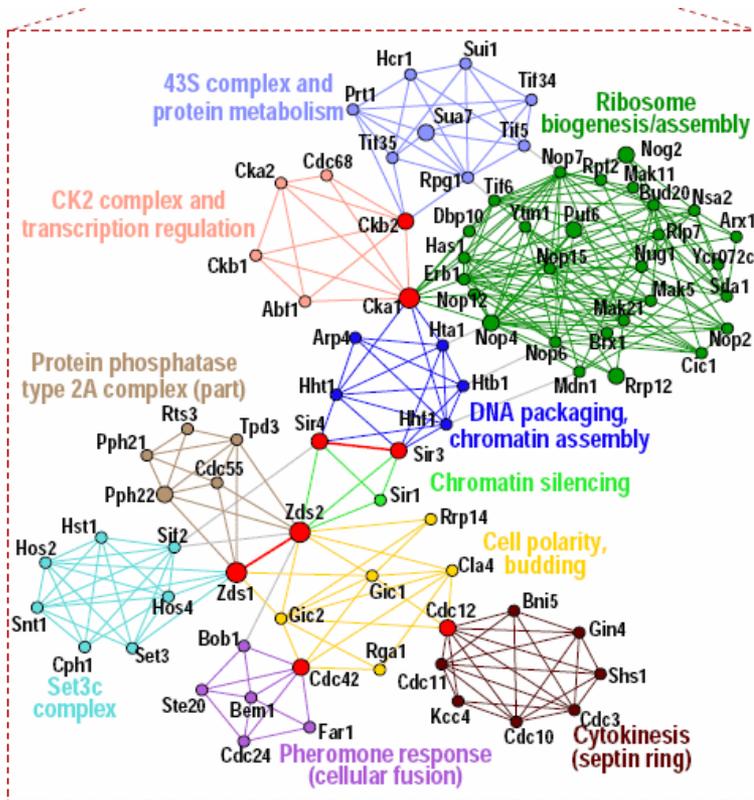
Análisis estadístico

- Clustering de datos
- Overlaps y correlaciones en datasets
- Enriquecimiento funcional

Modelado

- Inferencia estadística de interacciones a partir de datos
- Modelado de sistemas dinámicos y simulaciones biofísicas

Enriquecimiento funcional



- **MsigDB**
- **Gene Ontology**
- **DAVID**



Molecular Signatures DataBase



- Conjunto de **bolsas de genes** armadas bajo algún criterio relevante desde el punto de vista biológico.
- Conocimiento biológico se encuentra embebido en la composición de cada conjunto

The MSigDB gene sets are divided into 8 major collections:

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

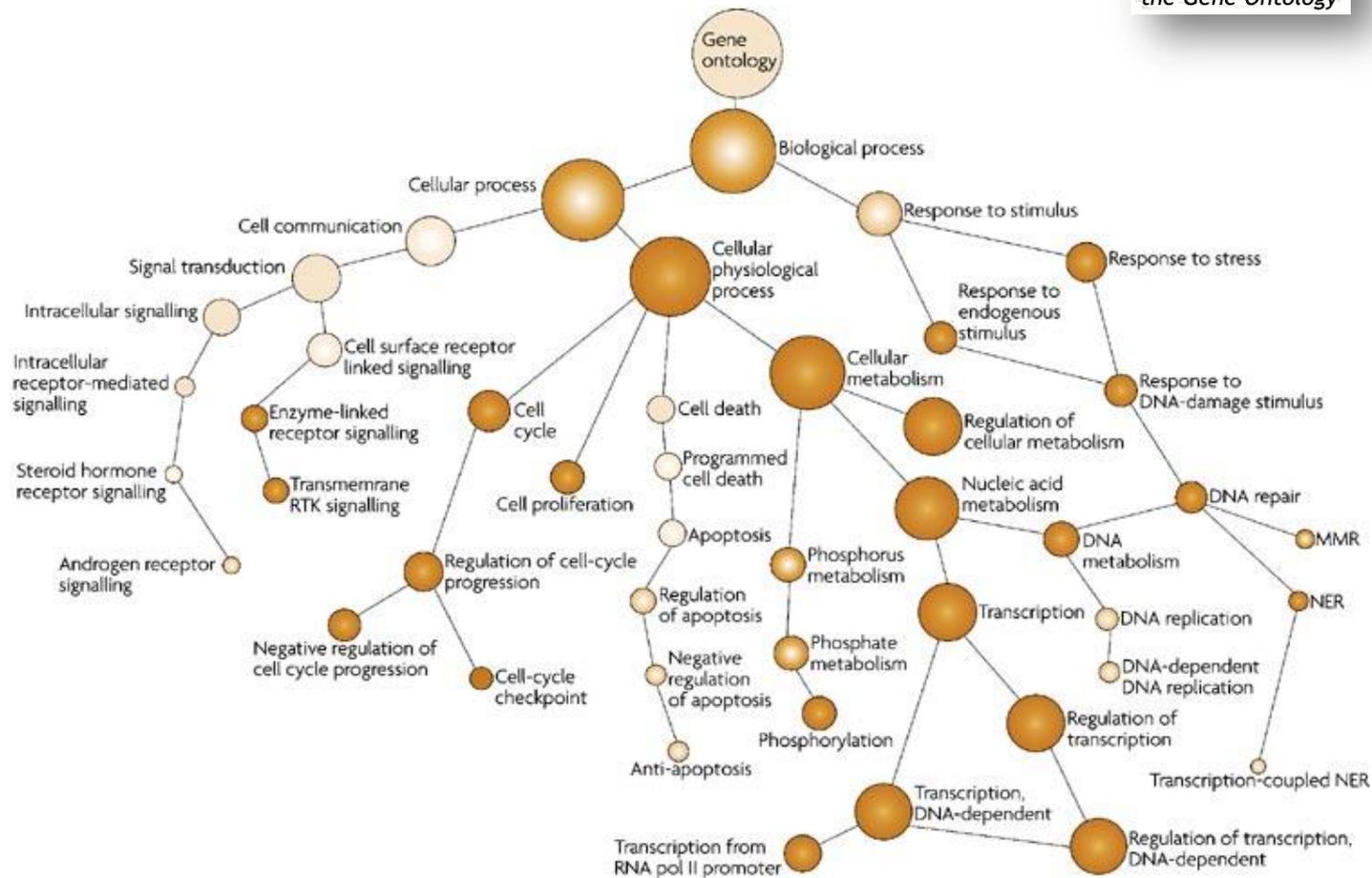
C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

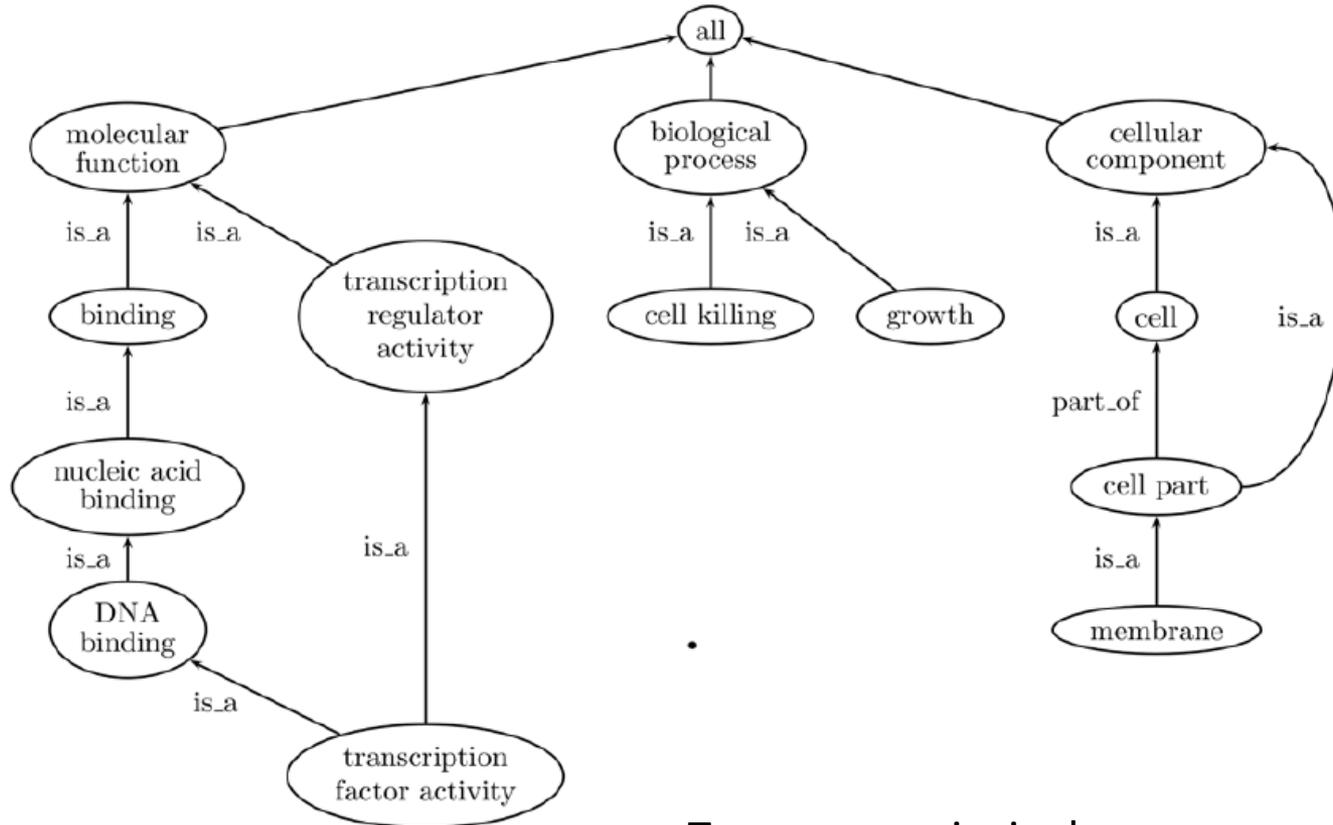
C7 **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

Gene Ontology



Nature Reviews | Cancer

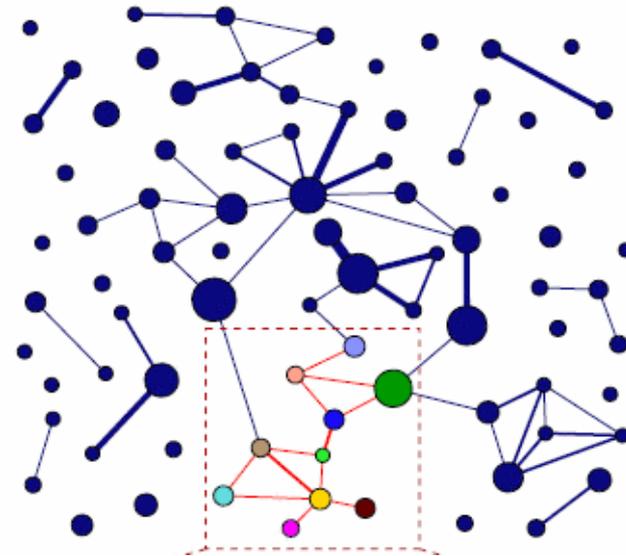
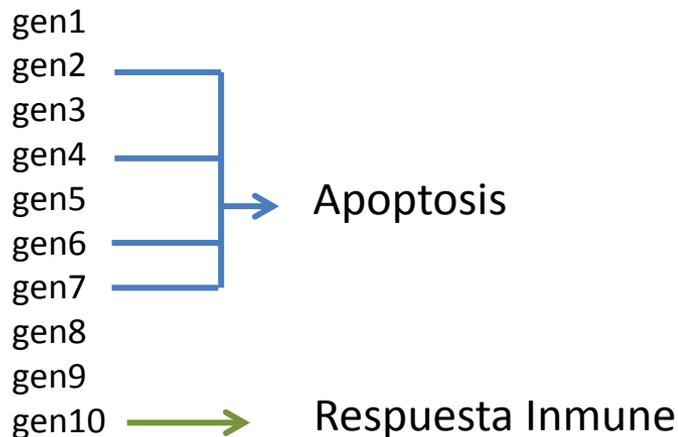
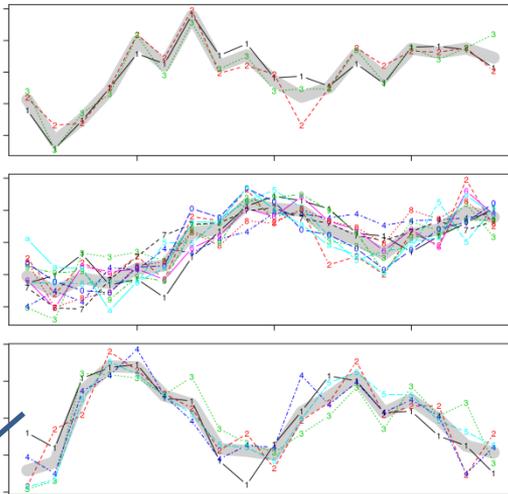
- Vocabulario controlado, organizado como DAG
- Organización cuasi-jerárquica – Especificidad de conceptos
- Productos génicos son mapeados a un (o más) nodos específicos y automáticamente son también anotados en nodos ancestros de la ontología



- Tres ramas principales
 - **Molecular Function** (Qué hace?)
 - **Cellular Component** (Dónde lo hace?)
 - **Biological Process** (Por qué lo hace?)

Enriquecimiento funcional

Algún concepto de **GO** o conjunto de **MSigDB** esta sobrerrepresentado en alguna comunidad o cluster detectado en mis datos?



Existe un **vinculo** entre pertenecer a una **comunidad** dada y estar anotado dentro de un **tema biologico**?

Test de Fisher Exacto

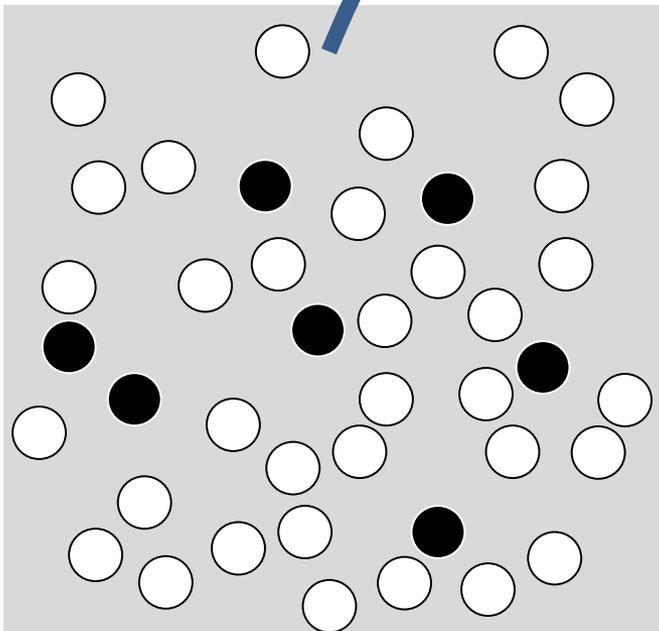
Que tan probable es observar un dado numero de hits en la lista solo por azar

Enriquecimiento funcional

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10



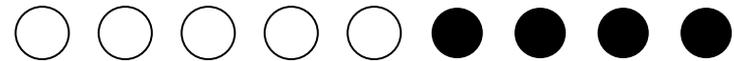
Apoptosis (r=12)



N (100) genes examinados

Test de Fisher Exacto

Que tan probable es observar un dado numero de hits en la lista solo por azar



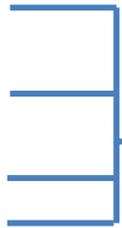
k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica

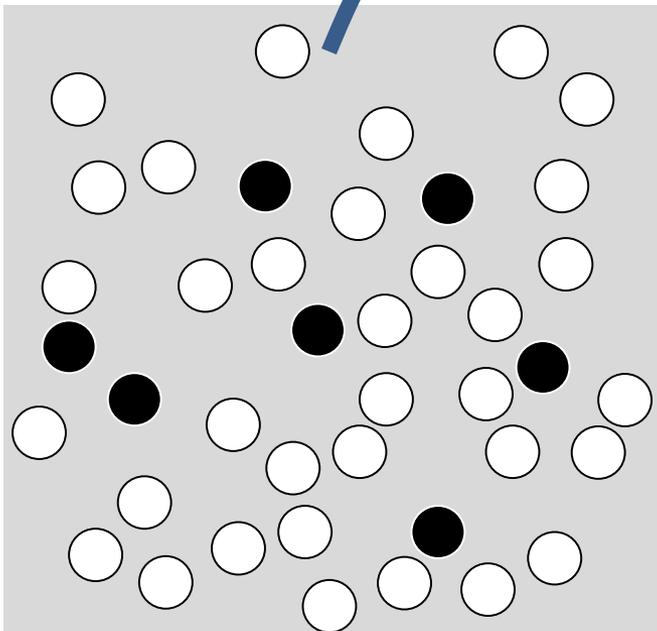
	categoria bio	categoria bio	
cluster	4	6	10
<u>cluster</u>	8	82	90
	12	88	100
	categoria bio	categoria bio	
cluster	m	k-m	k
<u>cluster</u>	r-m	N-k+m-r	N-k
	r	N-r	N

Enriquecimiento funcional

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10



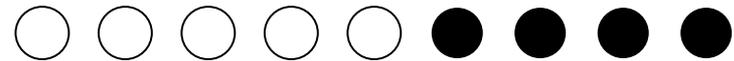
Apoptosis ($r=12$)



N (100) genes examinados

Test de Fisher Exacto

Que tan probable es observar un dado numero de hits en la lista solo por azar



k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica

	categoria bio	<u>categoria bio</u>	
cluster	m	$k-m$	k
<u>cluster</u>	$r-m$	$N-k+m-r$	$N-k$
	r	$N-r$	N

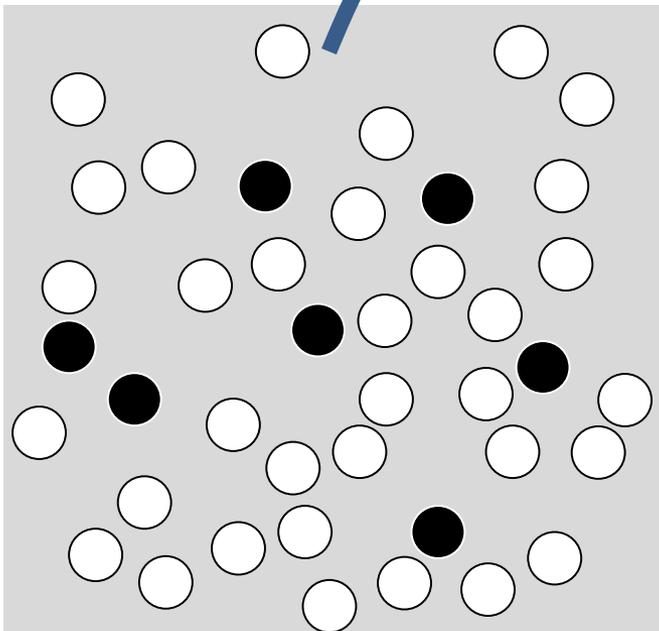
Cual es la probabilidad de ver esta tabla o alguna **mas extrema** si saco las bolas al azar? (i.e. no hay coneccion alguna entre pertenencia al cluster y a la categoria biológica?)

Enriquecimiento funcional

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10



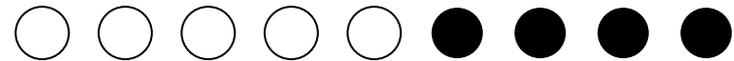
Apoptosis (r=12)



N (100) genes examinados

Test de Fisher Exacto

Que tan probable es observar un dado numero de hits en la lista solo por azar



k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica

	categoria bio	<u>categoria bio</u>	
cluster	m	k-m	k
<u>cluster</u>	r-m	N-k+m-r	N-k
	r	N-r	N

Cual es la probabilidad de ver esta tabla?

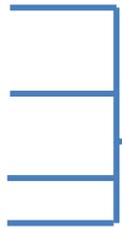
$$p = \frac{\binom{k}{m} \binom{N-k}{r-m}}{\binom{N}{r}}$$

← maneras de elegir r bolas negras con m afuera (en un grupo de k)

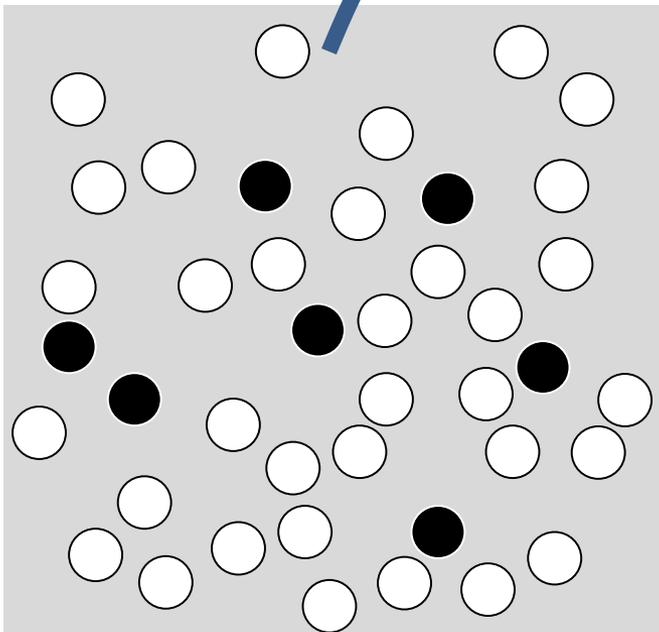
← maneras de elegir r bolas negras sin importar si es adentro o afuera

Enriquecimiento funcional

gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10



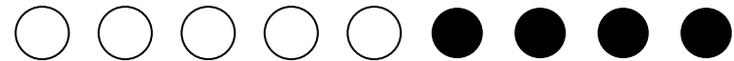
Apoptosis (r=12)



N (100) genes examinados

Test de Fisher Exacto

Que tan probable es observar un dado numero de hits en la lista solo por azar



k (10) genes en la comunidad/cluster

m (4) genes del cluster en la categoria biológica

	categoria bio	categoria bio	
cluster	m	k-m	k
cluster	r-m	N-k+m-r	N-k
	r	N-r	N

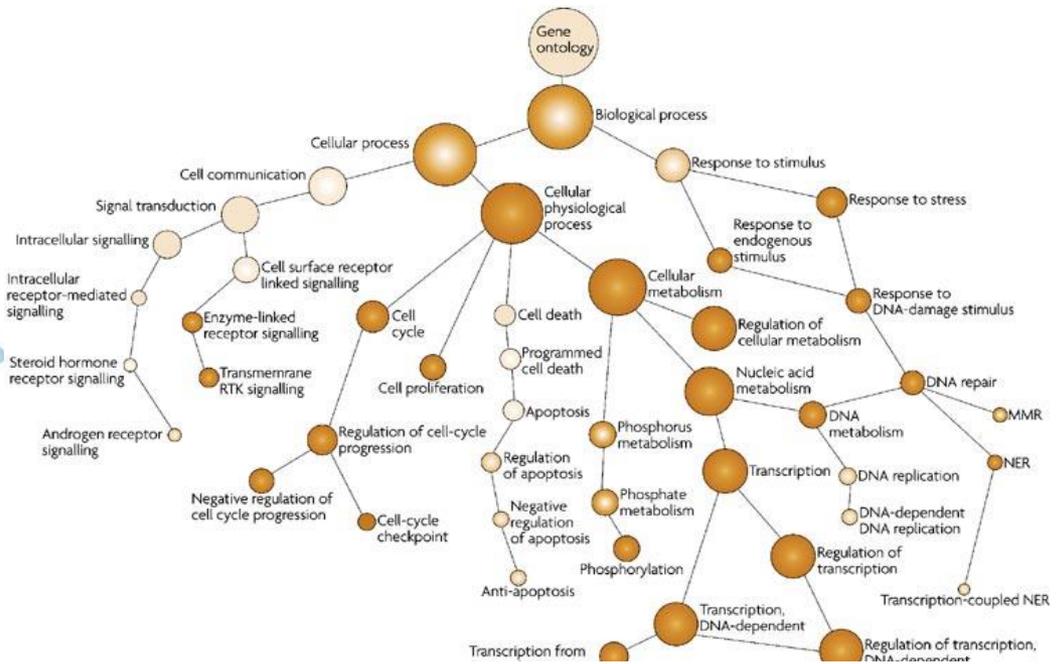
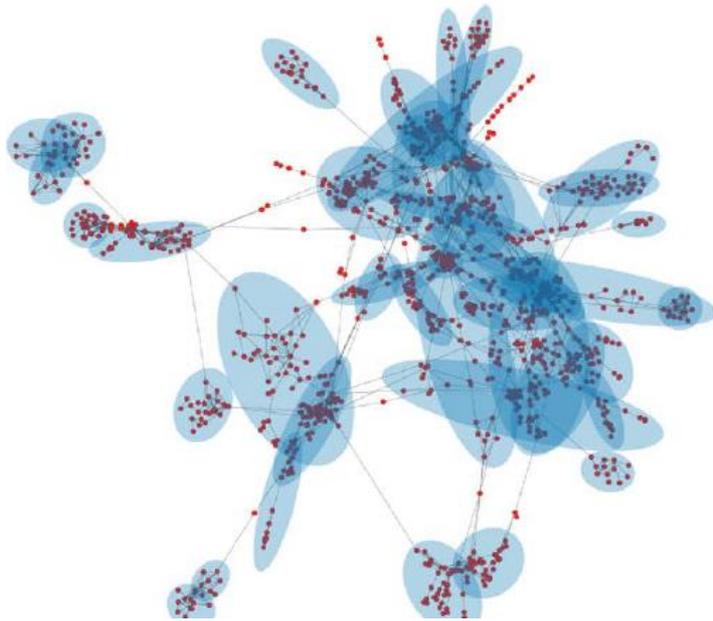
Cual es la probabilidad de ver esta tabla u **otra más extrema?**

$$p_{valor} = \sum_{m'=m}^k \frac{\binom{k}{m'} \binom{N-k}{r-m'}}{\binom{N}{r}}$$

Es chica esta cantidad?

$$p_v < p_{umbral}$$

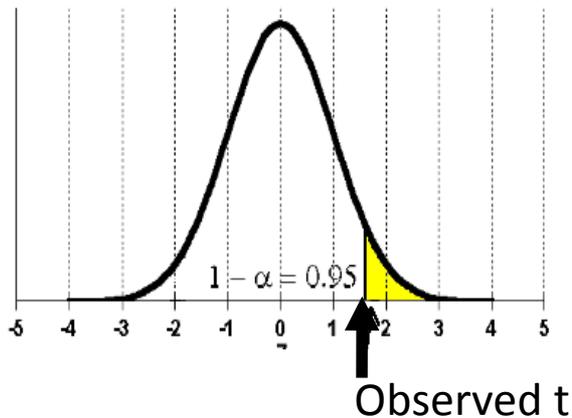
El problema de los grandes datos: testeo de múltiples hipótesis



- Para una dada comunidad/cluster voy a realizar un **test sobre TODO posible concepto GO** para buscar sobrerrepresentación.....son muchos tests (!)
- Existe la posibilidad de **ver eventos extremos** en el **conjunto de todos los test individuales** realizados por azar, aún cuando no haya vínculo entre cluster y anotación
- Eventos con $p_v=0.05$ tienen un 5% de prob de ocurrir por azar Entonces espero ~ 5 Falsos Positivos si hago 100 tests y uso 0.05 como corte de significancia de test s individuales.

Ejemplo del problema de testeo múltiple de hipótesis

Distribution for “t” under Null Hypothesis



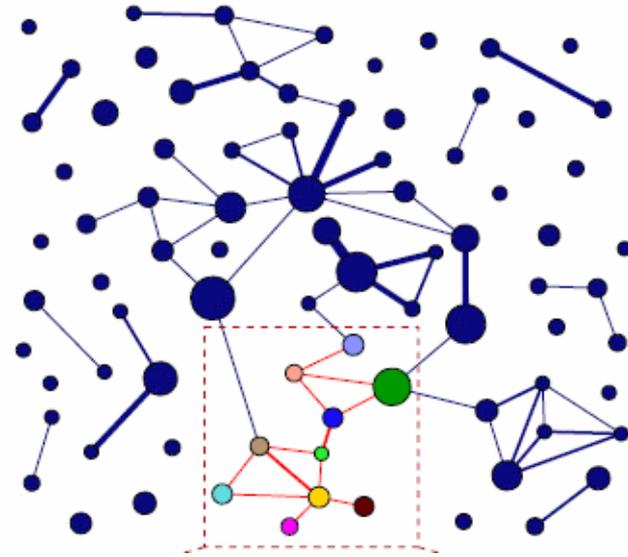
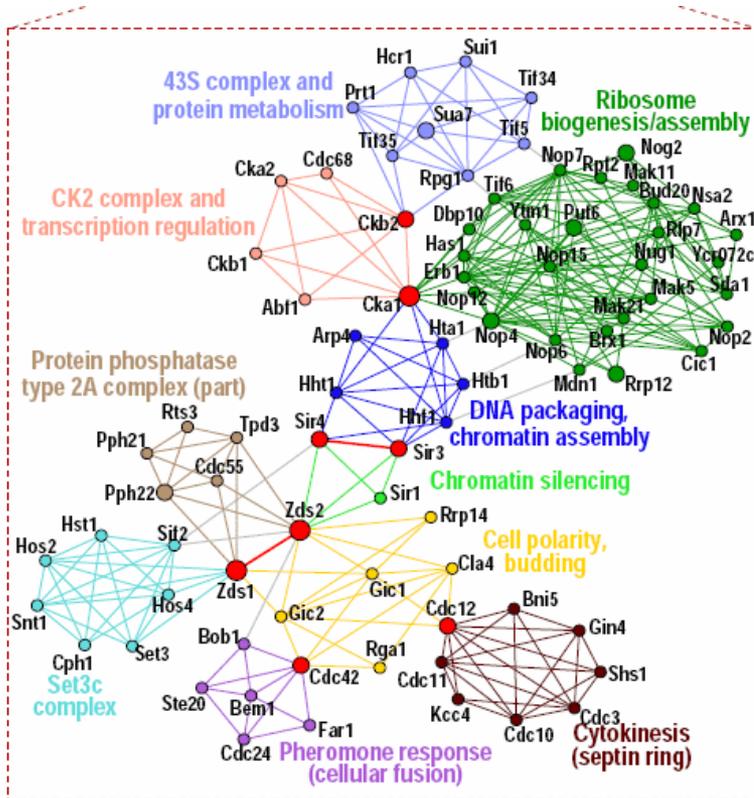
Random generated “expression matrix”

	A				B				Test p-value
1	0,15	0,51	0,08	0,84	0,17	0,91	0,69	0,37	0,58
2	0,47	0,06	0,28	0,12	0,83	0,95	0,42	0,29	0,08
3	0,07	0,94	0,17	0,59	0,66	0,1	0,21	0,2	0,55
4	0,17	0,49	0,88	0,05	0,33	0,11	0,7	0,7	0,8
5	0,86	0,4	0,6	0,26	0,57	0,48	0,7	0,28	0,87
6	0,22	0,04	0,41	0,7	0,29	0,4	0,5	0,44	0,68
7	0,05	0,45	0,06	0	0,96	0,3	0,53	0,45	0,06
8	0,49	0,45	0,55	0,25	0,59	0,83	0,28	0,89	0,22
9	0,83	0,51	0,26	0,64	0,83	0,46	0,63	0,79	0,46
10	0,03	0,33	0,6	0,86	0,09	0,78	0,02	0,96	0,99
11	0,72	0,81	0,54	0,14	0,64	1	0,22	0,82	0,61
12	0,44	0,16	0,21	0,58	0,6	0,89	0,85	0,64	0,02
13	0,5	0,56	0,11	0,75	0,55	0,91	0,7	0,88	0,13
14	0,48	0,17	0,07	0,3	0,83	0,16	0,11	0,87	0,33
15	0,03	0,84	0,35	0,26	0,45	0,71	0,03	0,96	0,55
16	0,38	0,11	0,02	0,58	0,15	0,93	0,23	0,51	0,43
17	0,39	0,56	0,12	0,48	0,02	0,93	0,95	0,89	0,25
18	0,58	0,57	0,22	0,4	0,03	0,76	0,81	0,57	0,63
19	0,82	0,68	0,04	0,63	0,91	0,83	0,88	0,75	0,14
20	0,84	0,77	0,64	0,89	0,72	0,23	0,11	0,91	0,19

False Positive at p_cutoff=0.05 !!!

Need of **Multiple Hypothesis Testing Correction** when testing ~10000 genes

Enriquecimiento funcional

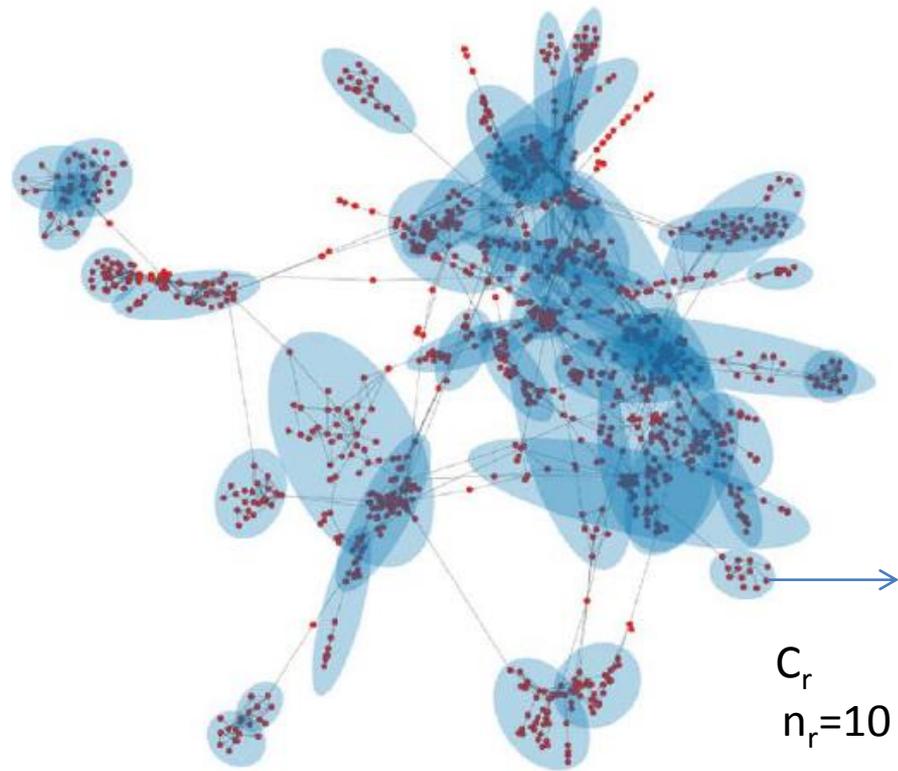
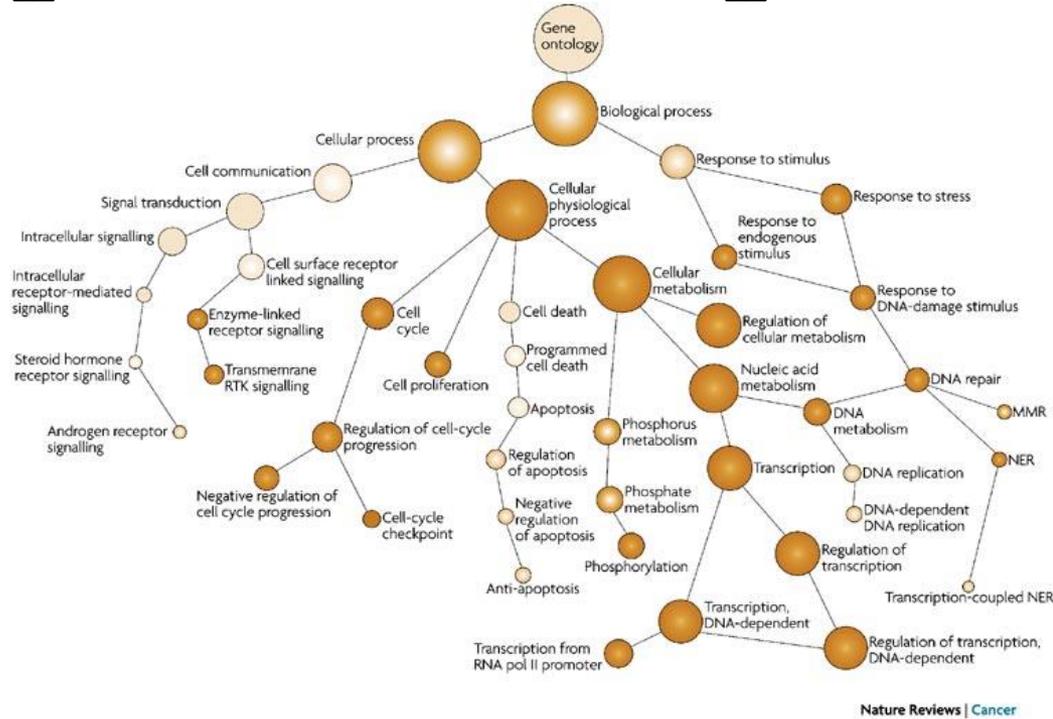


- MsigDB
- Gene Ontology
- DAVID



Indice de Homogeneidad Biológica

Validación de comunidades utilizando conocimiento externo



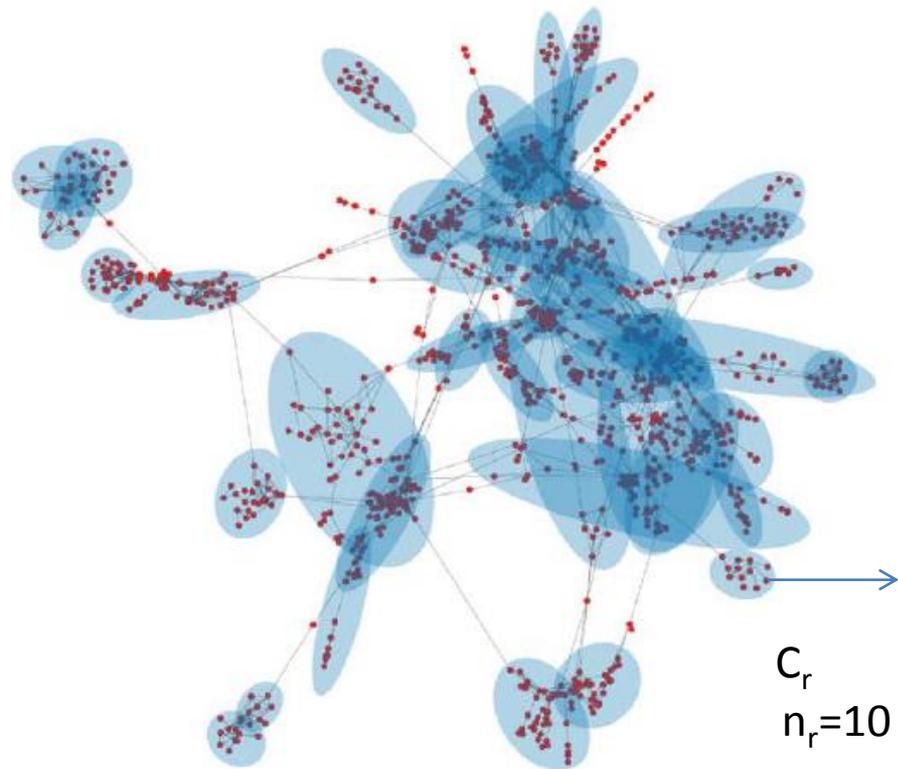
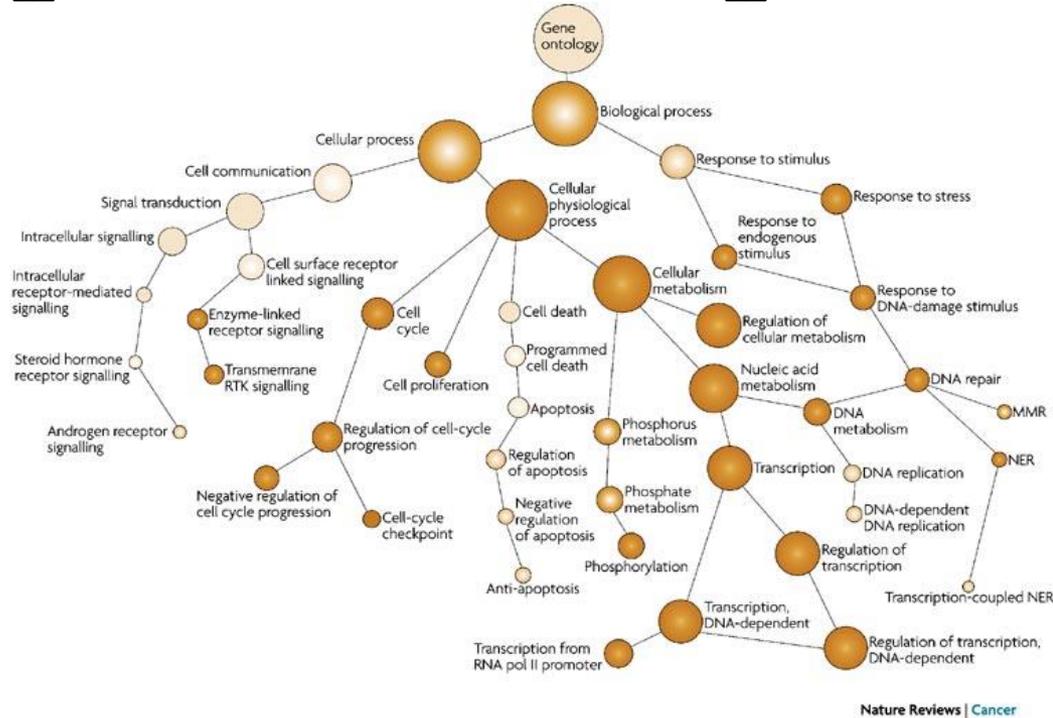
$$BHI_r = \frac{2}{n_r(n_r - 1)} \sum_{i \neq j \in C_r} \delta_{GO_i, GO_j}$$

$$BHI = \frac{1}{k} \sum_{r=1}^k BHI_r$$

↑
criterio binario de coincidencia en espacio GO

Índice de Homogeneidad Biológica

Validación de comunidades utilizando conocimiento externo



gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

C_r
 $n_r=10$

Así como es posible inferir una métrica a partir de la red también es posible inferir una **métrica** a partir de la estructura de **GO** que permita cuantificar el nivel de **similitud** entre proteínas en un **espacio de conocimiento biológico**.

Hacia una métrica GO

Cuanta información gano si me entero que una proteína está anotada en una clase GO_i?

Contenido de información

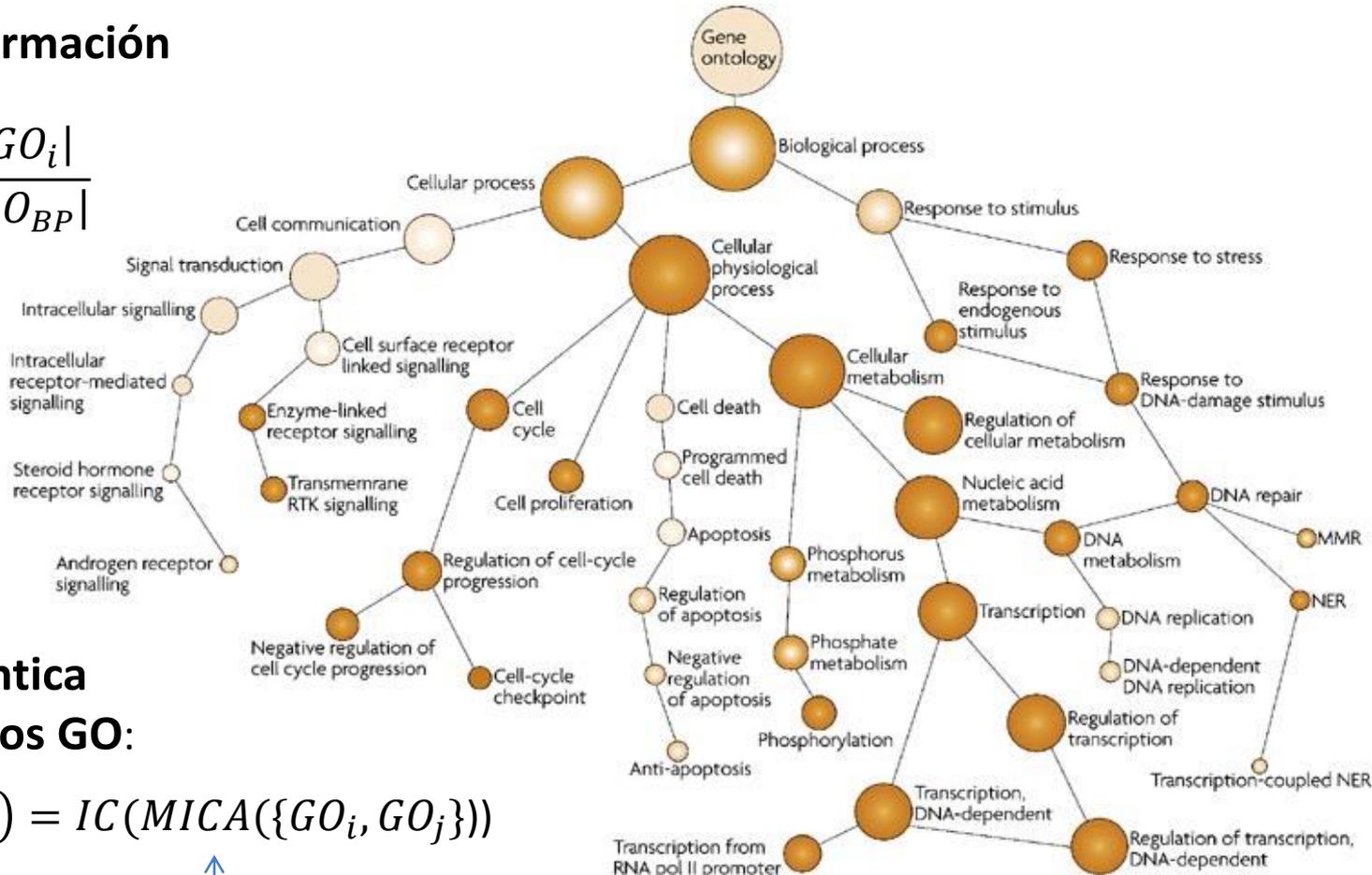
de un concepto GO:

$$IC(GO_i) = -\log \frac{|GO_i|}{|GO_{BP}|}$$

Similaridad semántica
entre dos conceptos GO:

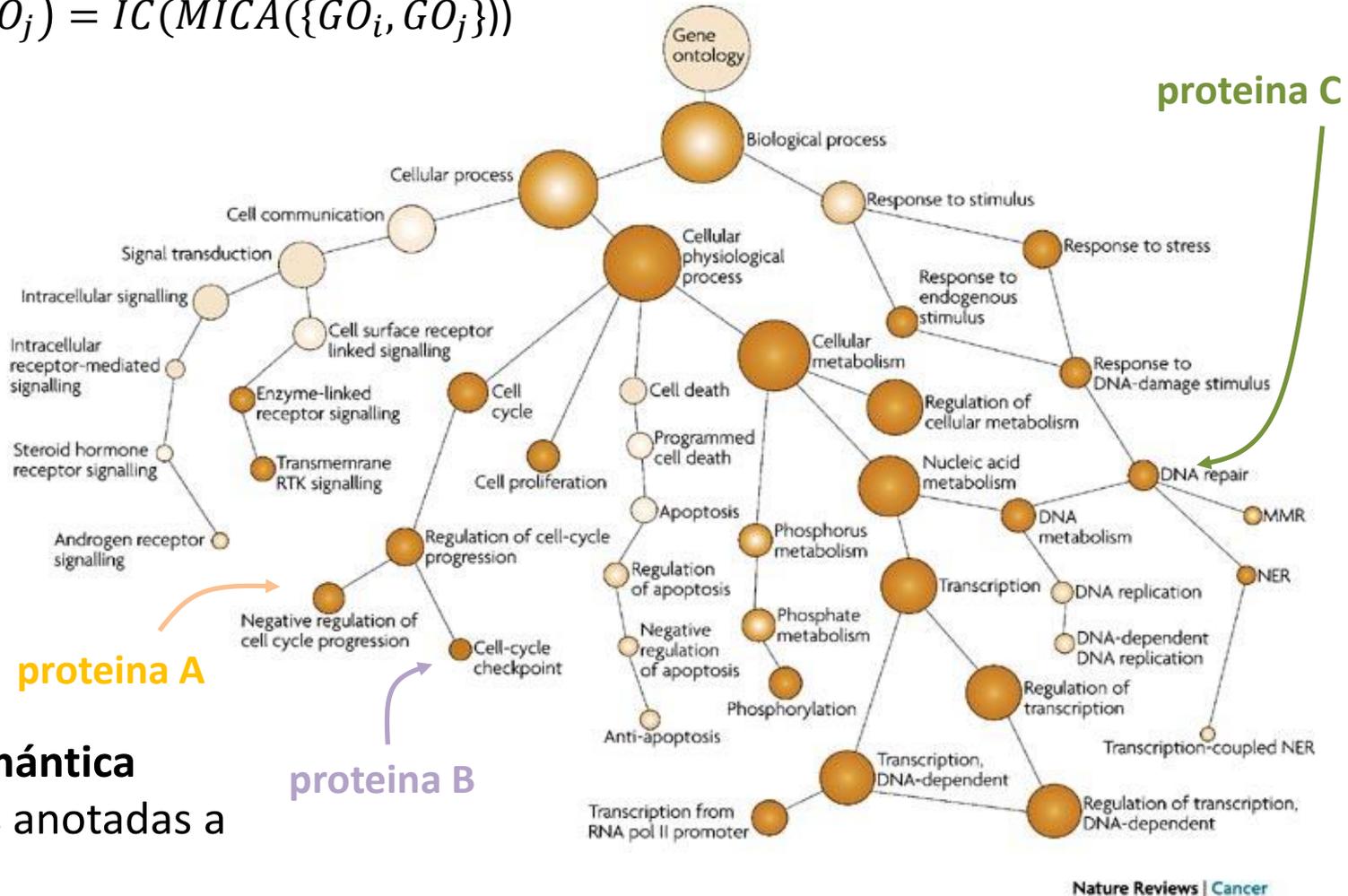
$$sim^{Resnik}(GO_i, GO_j) = IC(MICA(\{GO_i, GO_j\}))$$

Most Informative Common Ancestor



Similaridad semántica

$$sim^{Resnik}(GO_i, GO_j) = IC(MICA(\{GO_i, GO_j\}))$$



Similaridad semántica entre proteínas anotadas a conceptos GO.

$$sim(prot_A, prot_B) = f(\{GO_i\}_A, \{GO_i\}_B)$$

Similaridad semántica

Similaridad semántica entre **conceptos**

$$sim^{Resnik}(GO_i, GO_j) = IC(MICA(\{GO_i, GO_j\}))$$

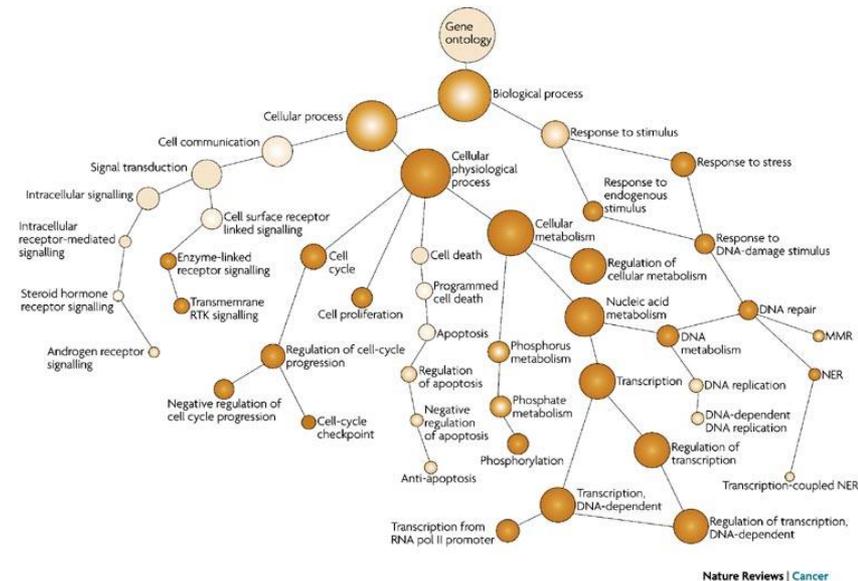
Similaridad semántica entre **proteínas**

$$prot_A \rightarrow \{GO_i\}_A \quad prot_B \rightarrow \{GO_i\}_B$$

GO_3 GO_5 GO_{223} GO_{542} GO_{23}

GO_1
 GO_{12}
 GO_{23}

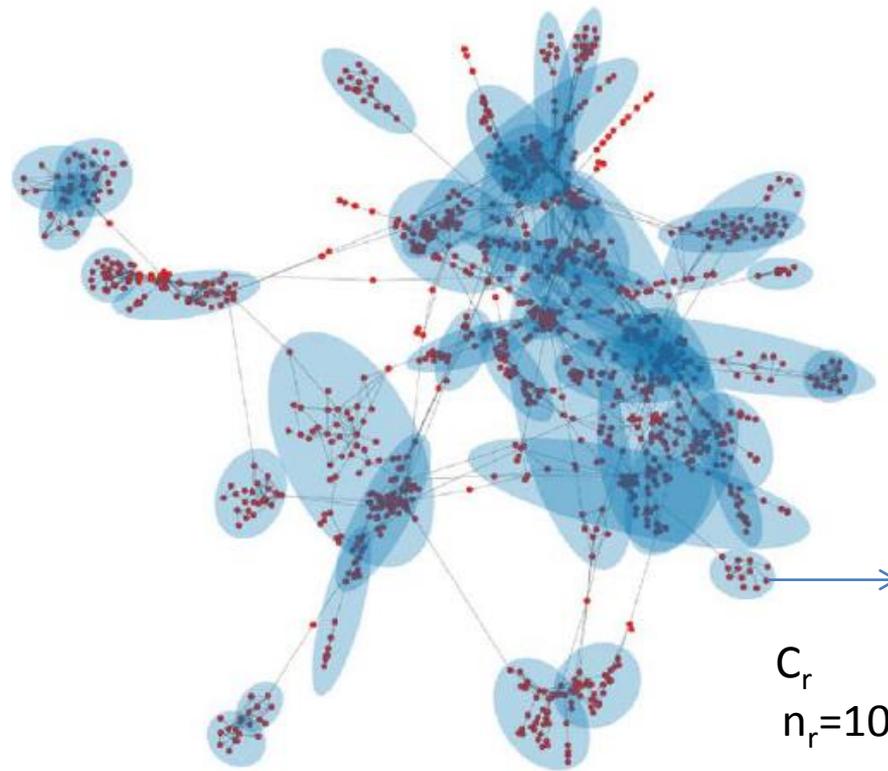
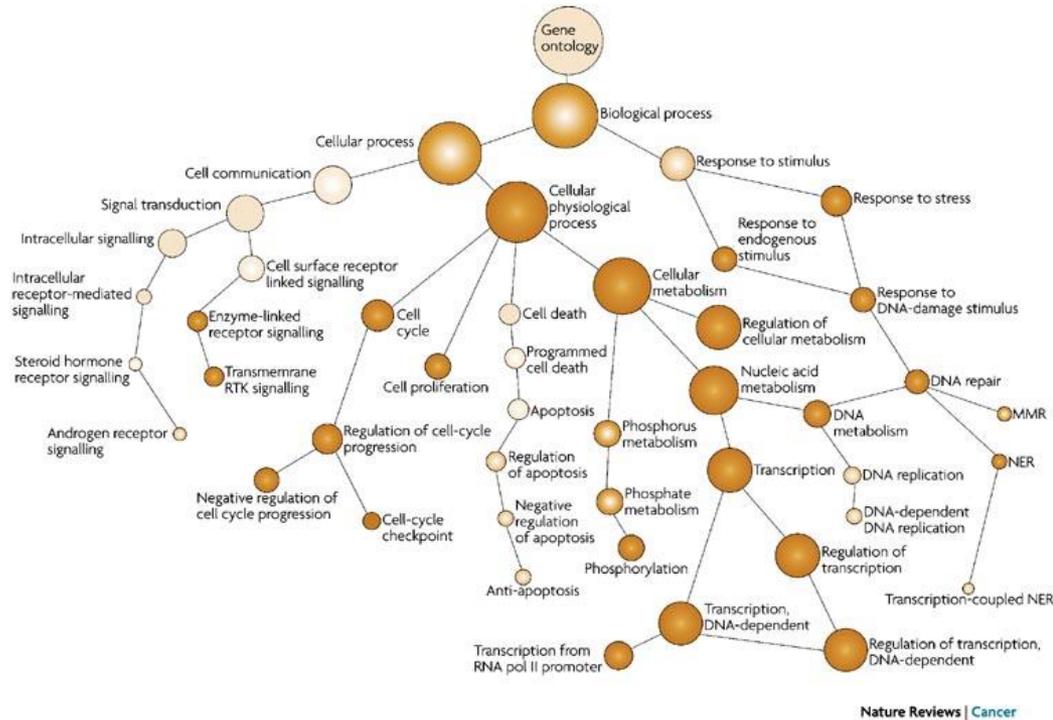
Matriz de sim GO's



$$sim(prot_A, prot_B) = f(\{GO_i\}_A, \{GO_i\}_B)$$

Indice de Homogeneidad Biológica Reloaded

Validación de comunidades utilizando conocimiento externo



gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

C_r
 $n_r=10$

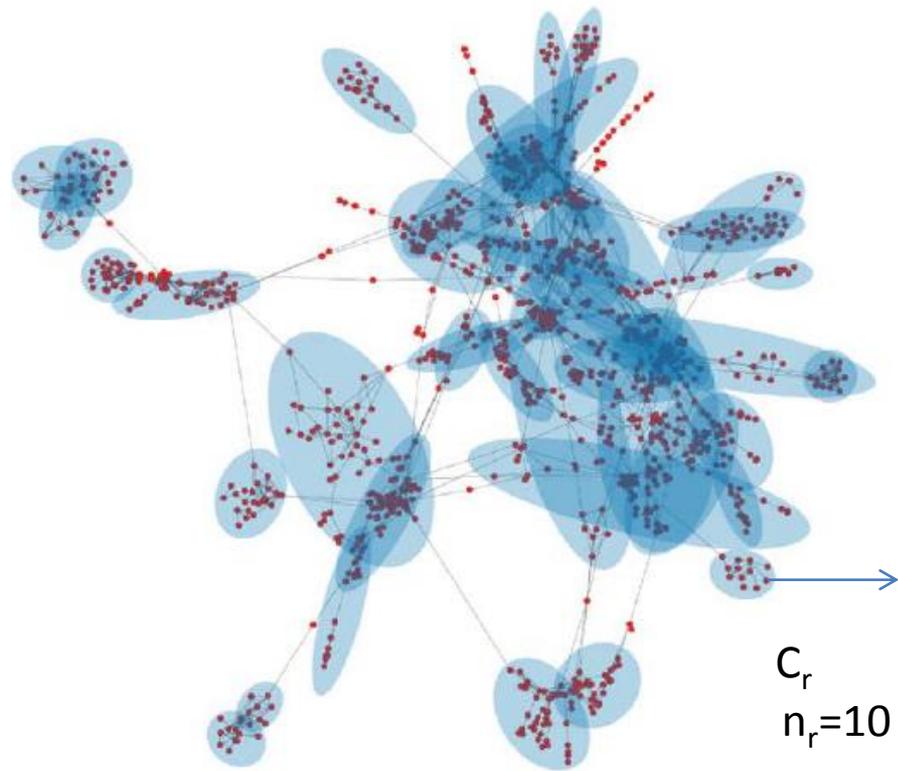
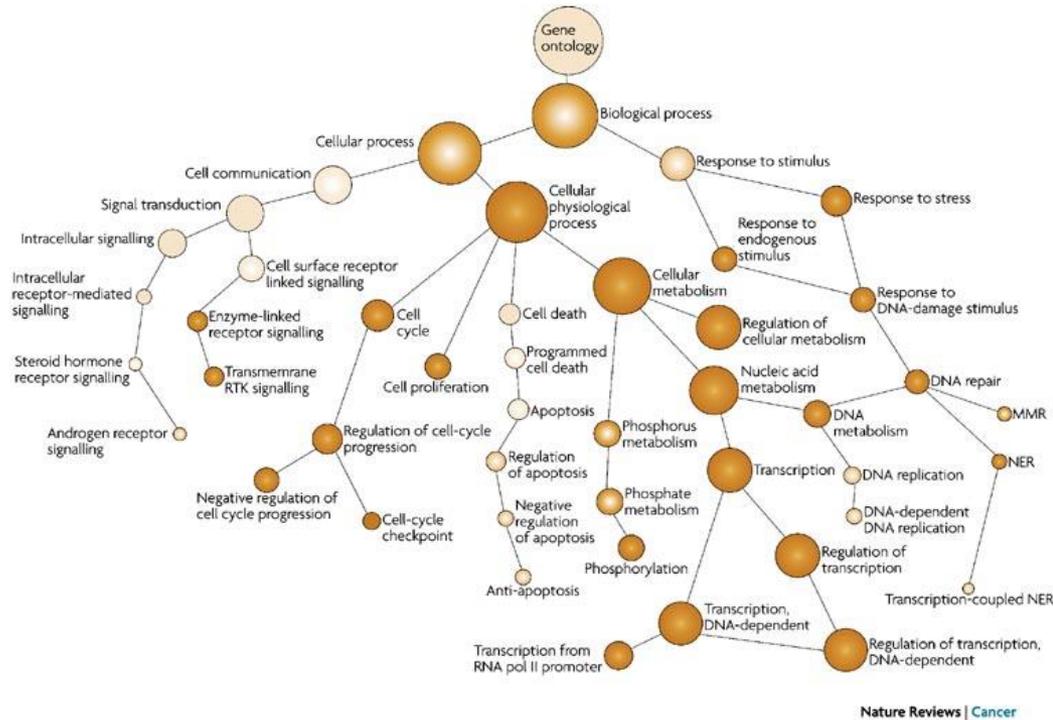
$$BHI_r = \frac{2}{n_r(n_r - 1)} \sum_{i \neq j \in C_r} \delta_{GO_i, GO_j}$$

$$BHI = \frac{1}{k} \sum_{r=1}^k BHI_r$$

↑
criterio binario de
coincidencia en
espacio GO

Indice de Homogeneidad Biológica Reloaded

Validación de comunidades utilizando conocimiento externo



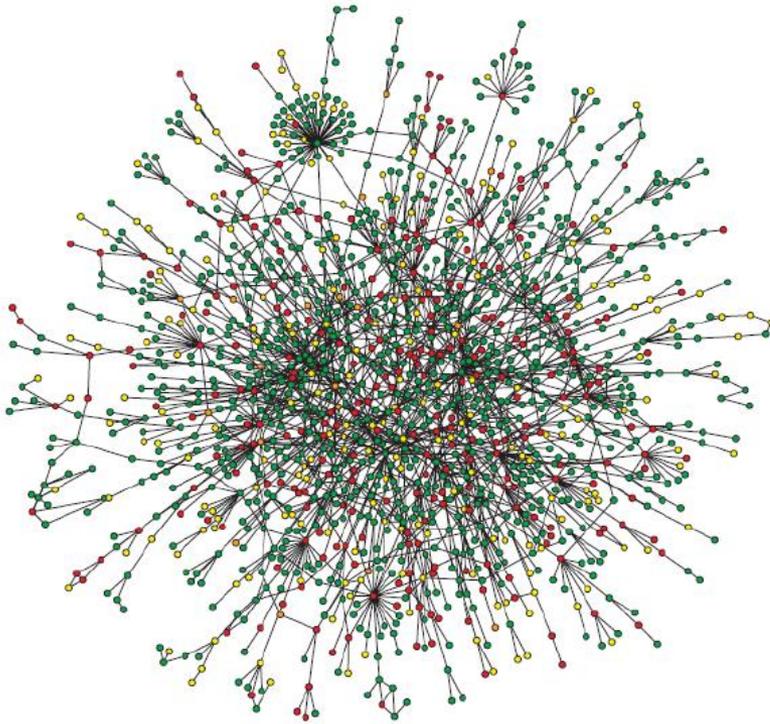
gen1
gen2
gen3
gen4
gen5
gen6
gen7
gen8
gen9
gen10

C_r
 $n_r=10$

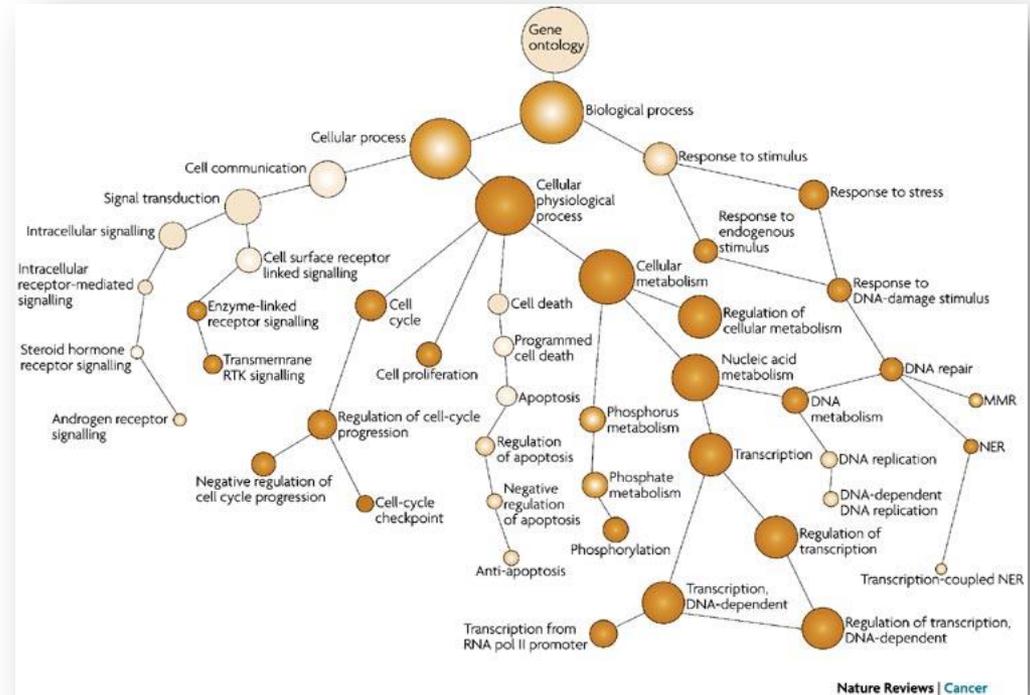
$$BHI_r = \frac{2}{n_r(n_r - 1)} \sum_{i \neq j \in C_r} sim_{GO}(i, j)$$

$$BHI = \frac{1}{k} \sum_{r=1}^k BHI_r$$

Integrando métricas



Gene Ontology



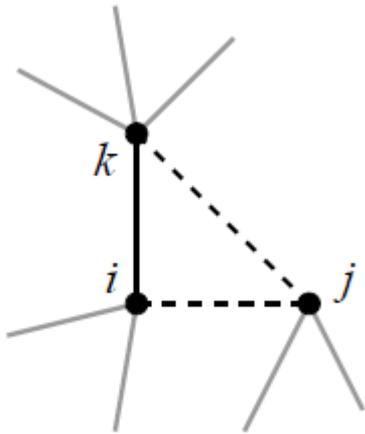
d_{red}

$d_{GO} = d_{semantic}$

Mixed Metrics

Buscamos estructuras **similares** en cuanto a la información embebida en la topología de la red **Y** el espacio de conocimiento biológico

Agrupando con métricas mixtas



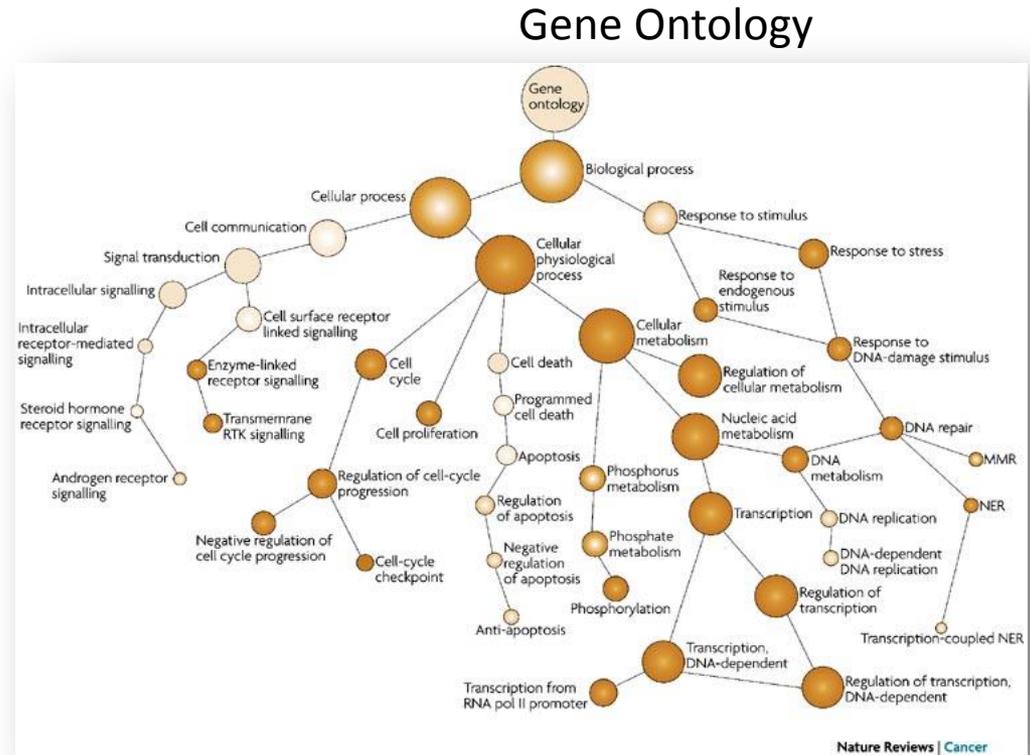
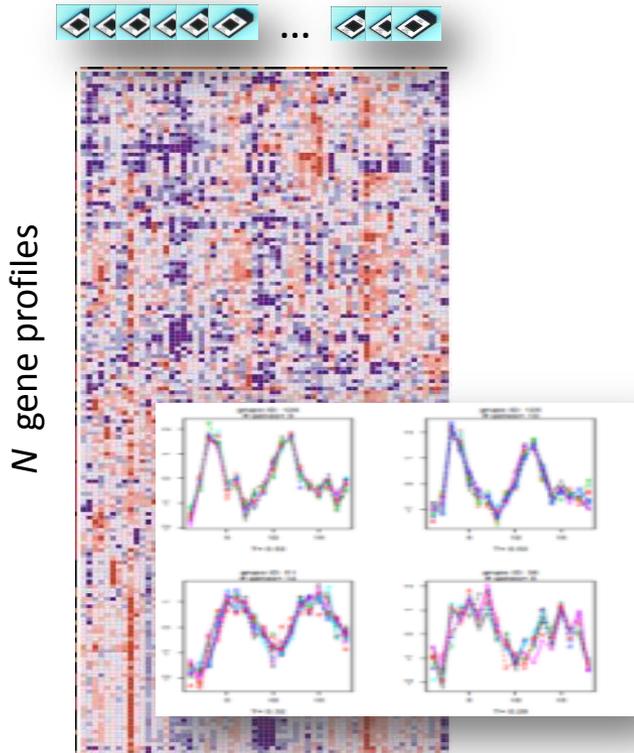
Equivalencia estructural **extendida**:

Dos vértices, i y j son similares si i tiene como vecino a k , quien a su vez es similar a j .

$$\sigma_{ij} = \alpha \sum_k A_{ik} \sigma_{kj} + \underline{s^{ext}}_{ij}$$

Nodos con alto grado tienen más vecinos, que pueden ser similares a terceros nodos, por lo que tienen más chances de aumentar su nivel de similaridad general. Si se desea mitigar este efecto...

Integrando métricas



$$dx = d_{\text{coexpression}}$$

$$d_{\text{GO}} = d_{\text{semantic}}$$

Mixed Metrics

Buscamos estructuras **similares** en cuanto a la información transcripcional **Y** el espacio de conocimiento biológico

Agrupando genes con métricas mixtas

El uso de métricas mixtas en algoritmos de reconocimiento de clusters / comunas permite buscar estructura simultáneamente coherente en ambos espacios (por ejemplo: espacio transcripcional y GO)

Las métricas pueden combinarse de diferente forma

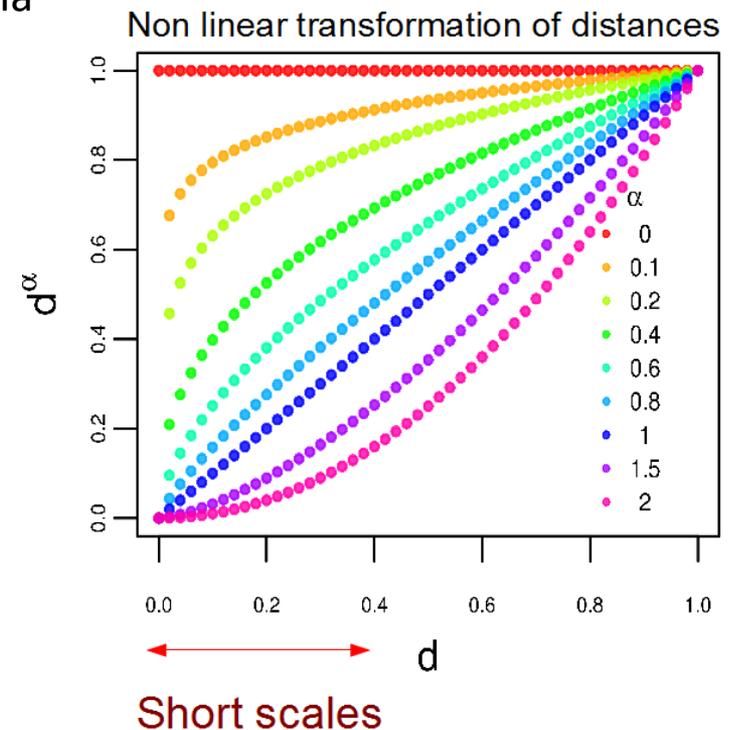
$$* d_{mix}^A = \sqrt{(1 - \alpha)d_x^2 + \alpha d_{go}^2}$$

$$* d_{mix}^B = d_x \cdot d_{go}^\alpha$$

$$* d_{mix}^C = d_x^{1-\alpha} \cdot d_{go}^\alpha$$

- El parámetro α controla la medida en la que el espacio GO es tenido en cuenta en la estimación de similitudes entre objetos.

- d^B **acerca** genes **biológicamente relacionados** en el sentido de GO

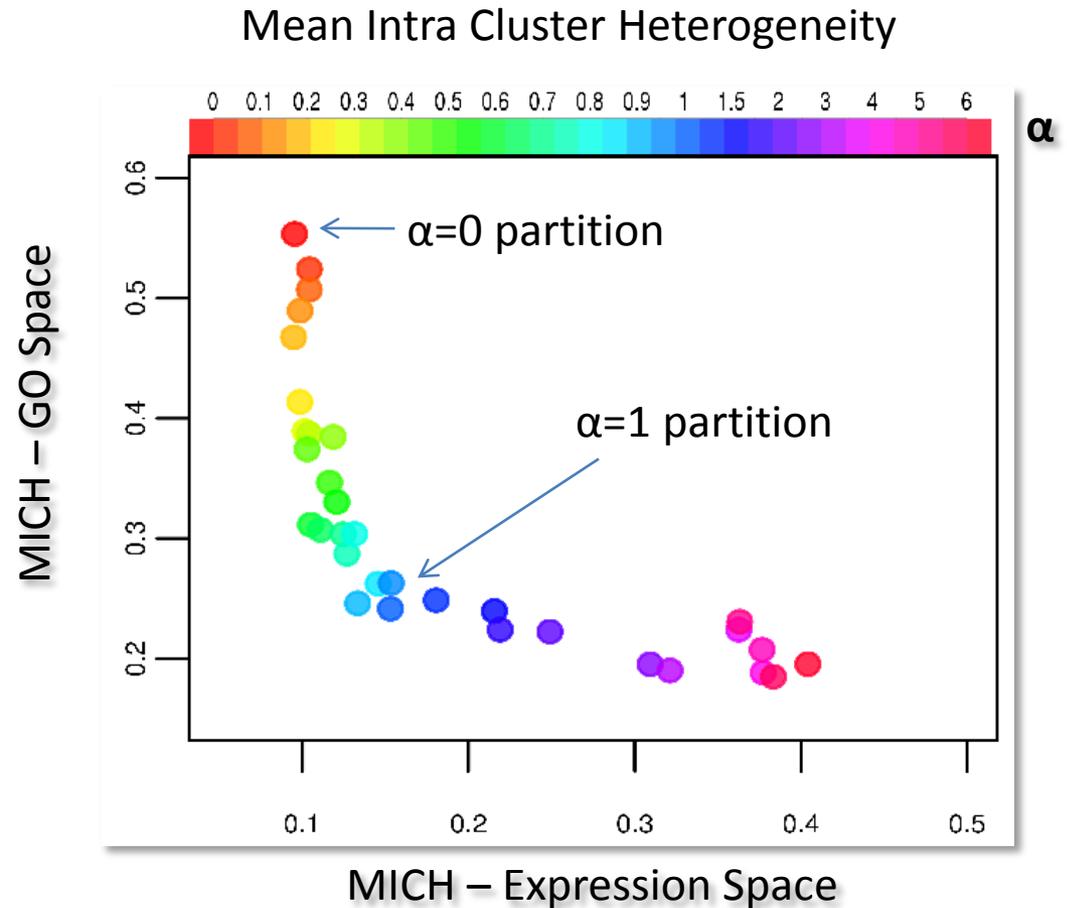
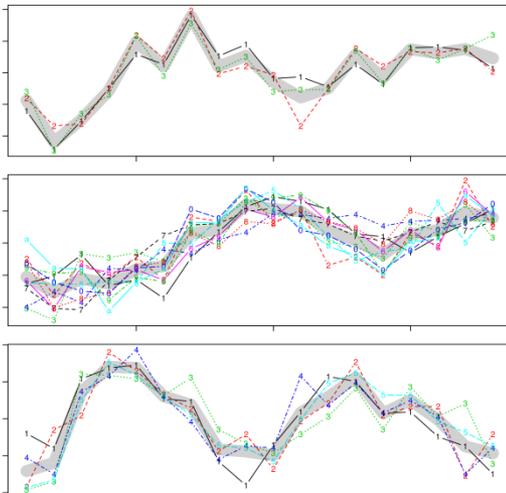


La transformación no-lineal altera de manera diferente la resolución a diferentes escalas

Agrupando genes con métricas mixtas

$$d_{mix} = d_x \cdot d_{go}^\alpha$$

Cluster structure



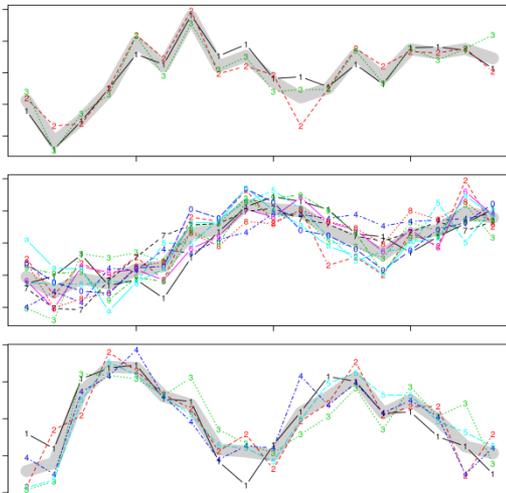
$$MICH(Y) = \frac{1}{N_{cl}} \sum_{cl_k}^{N_{cl}} \frac{2}{n_k(n_k-1)} \sum_{i < j \in cl_k}^{n_k} d_{ij}^Y$$

Agrupando genes con métricas mixtas

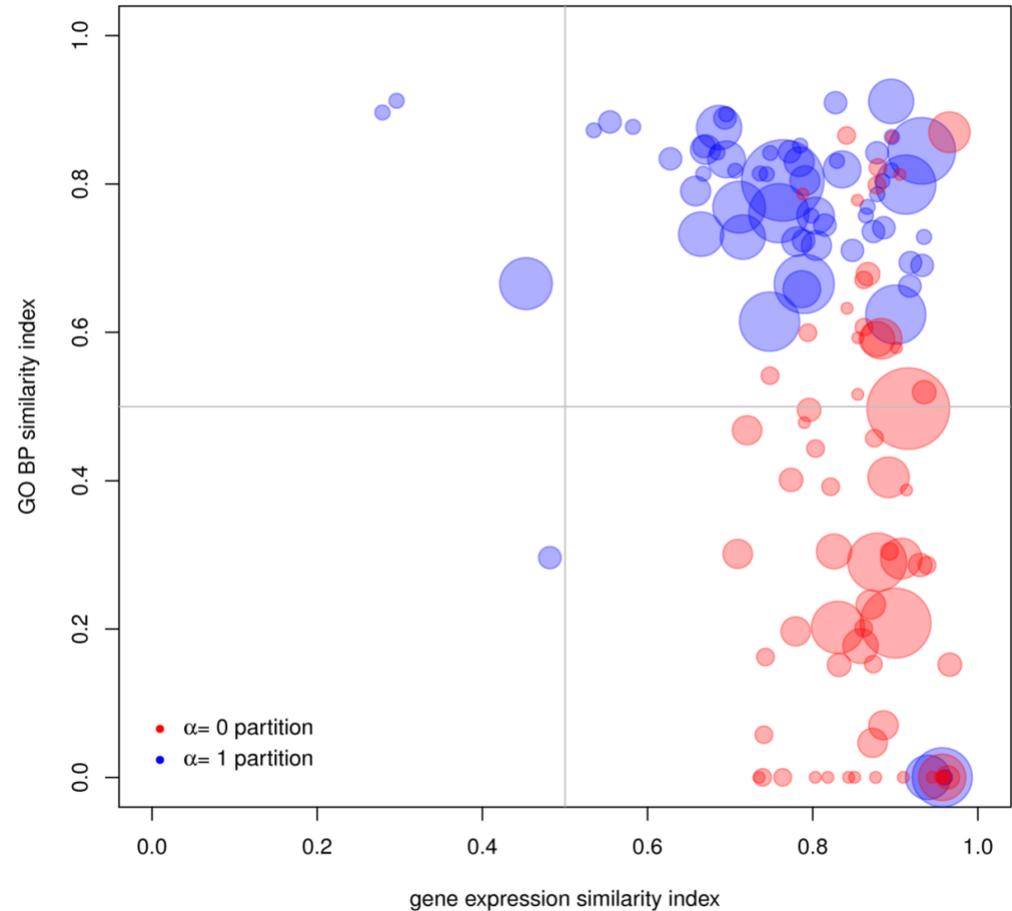
$$d_{mix} = d_x \cdot d_{go}^\alpha$$



Cluster structure



Clusters in similarity plane





SAMBA + CLICK