



Aplicación de diferentes métricas de dependencia entre variables en la detección de coexpresión de genes en el arroz



Nicolás Antonio López Rozo¹

Magíster en Ingeniería Civil
Estudiante de Doctorado en Ingeniería (ÓMICAS)
Pontificia Universidad Javeriana de Cali
30/08/2019

Results

Ranking

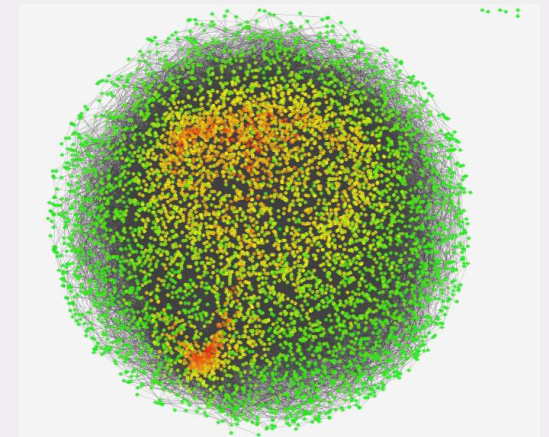
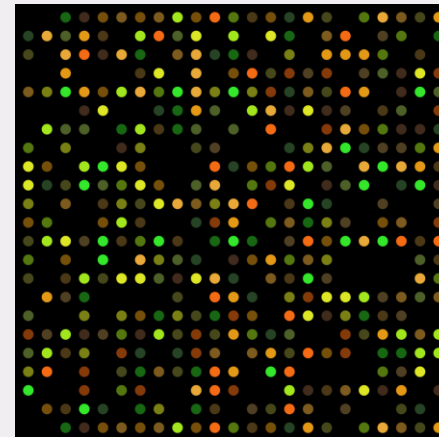
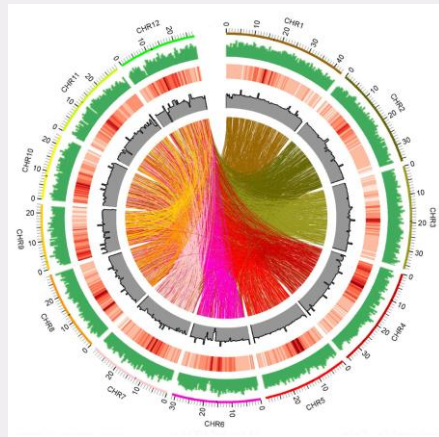
Metrics

Pearson

Problem

CONTENIDO

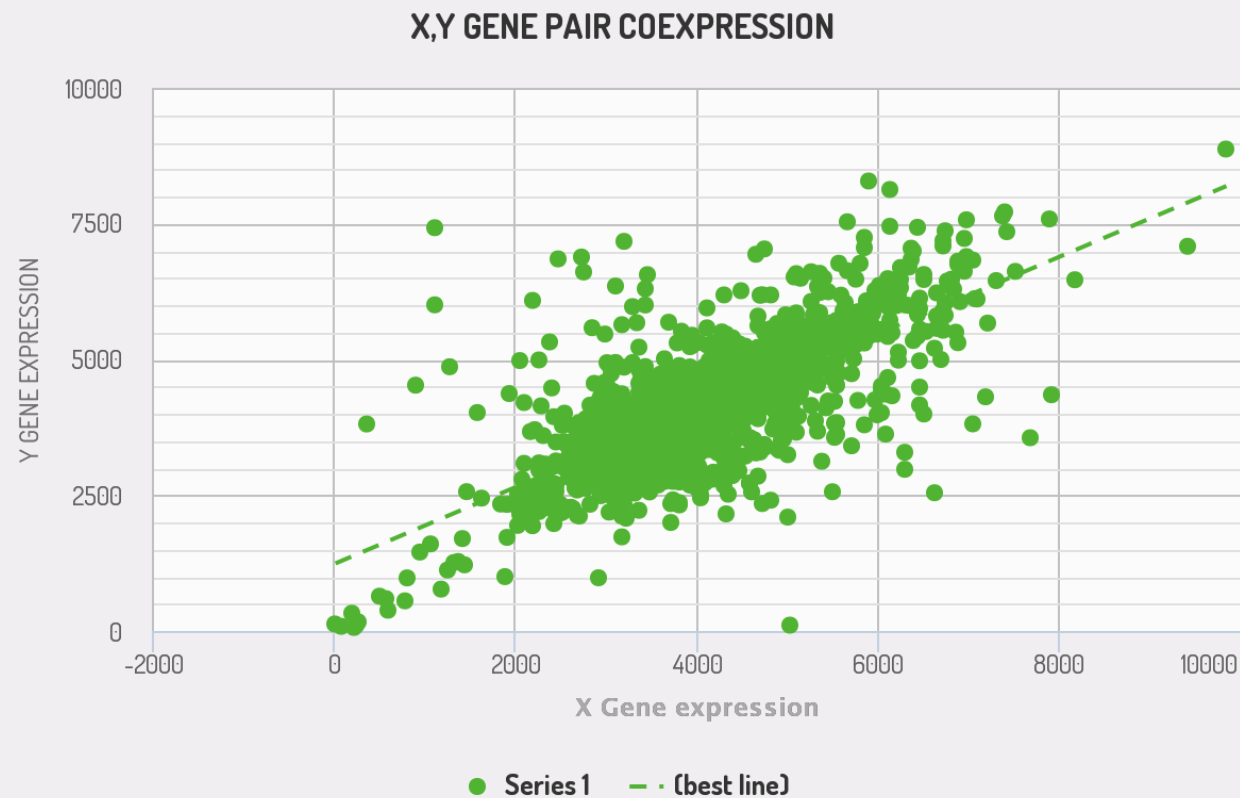
1. Introducción
2. Coeficiente de Correlación de Pearson
3. Maximal Information Coefficient (MIC)
4. Randomized Information Coefficient (RIC)
5. Distance Correlation (dCor)
6. Biweight Midcorrelation (BiCor)
7. Metodología de Mutual Rank
8. Resultados y Discusión



Contents

1. INTRODUCCIÓN

Dados los perfiles de expresión genética de un par de genes, si se grafica uno en términos del otro se puede obtener algo así:



meta-chart.com



1. INTRODUCCIÓN

Tradicionalmente, se utiliza el coeficiente de correlación de Pearson (PCC) para determinar la fuerza de esta correlación.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

No obstante, PCC no permite capturar adecuadamente relaciones no lineales (e.g. exponenciales) y por tanto, en este estudio se exploraron diferentes métricas que capturen otro tipo de relaciones no lineales.

- **Maximal Information Coefficient**
- **Randomized Information Coefficient**
- **Distance Correlation**
- **Biweight Midcorrelation**



2. COEF. DE PEARSON

Determina la fuerza con la que dos variables están linealmente relacionadas.

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}}$$

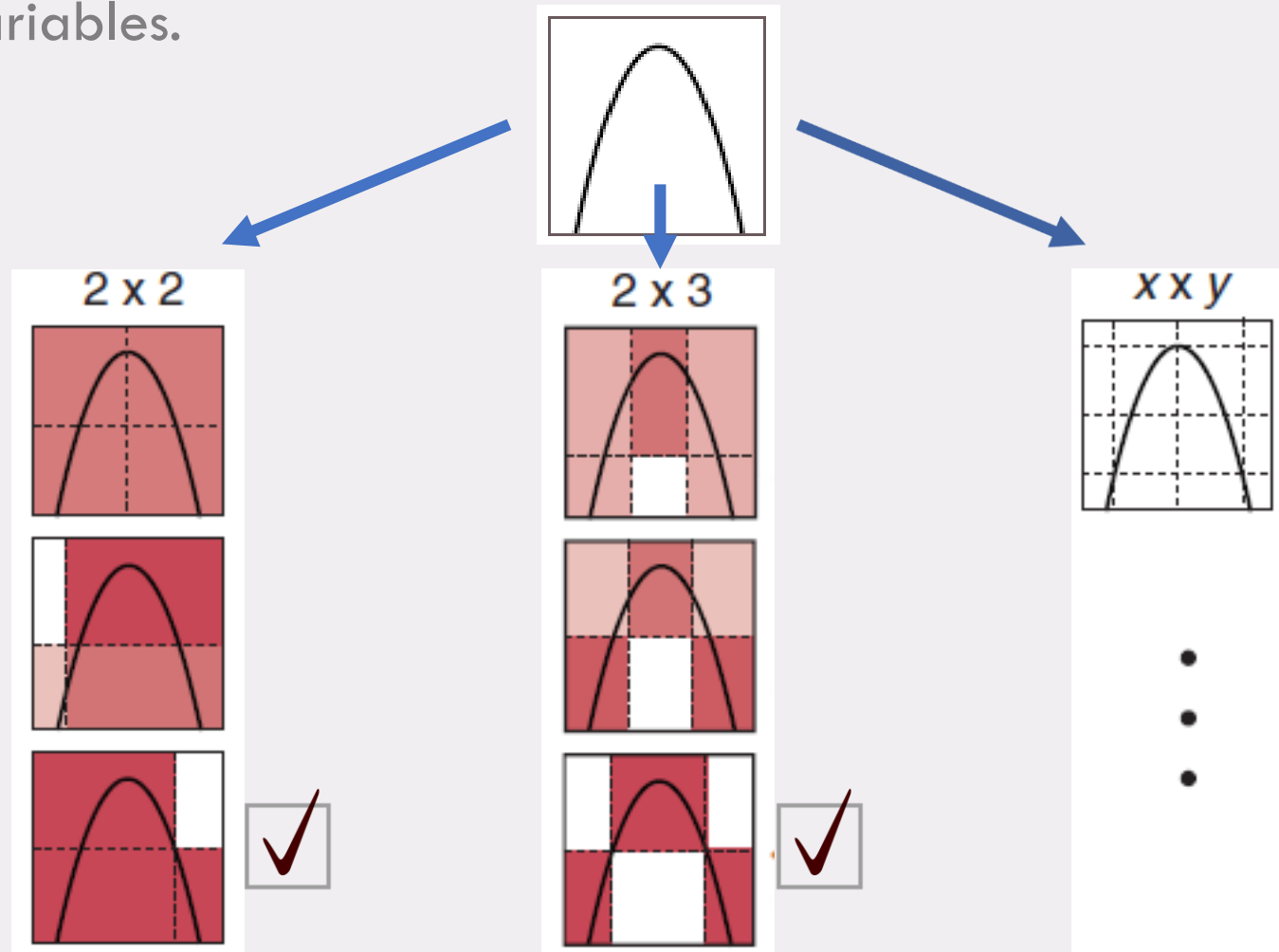
Existen ciertas guías para evaluar la fuerza de una correlación, las cuales dependen del campo de aplicación.

- $|r| < 0.1$: Correlación lineal despreciable
- $0.1 < |r| < 0.3$: Correlación lineal débil
- $0.3 < |r| < 0.5$: Correlación lineal moderada
- $|r| > 0.5$: Correlación lineal fuerte



3. MIC

El Maximal Information Coefficient (MIC) es una medida de la fuerza de una asociación lineal o no lineal entre dos variables.



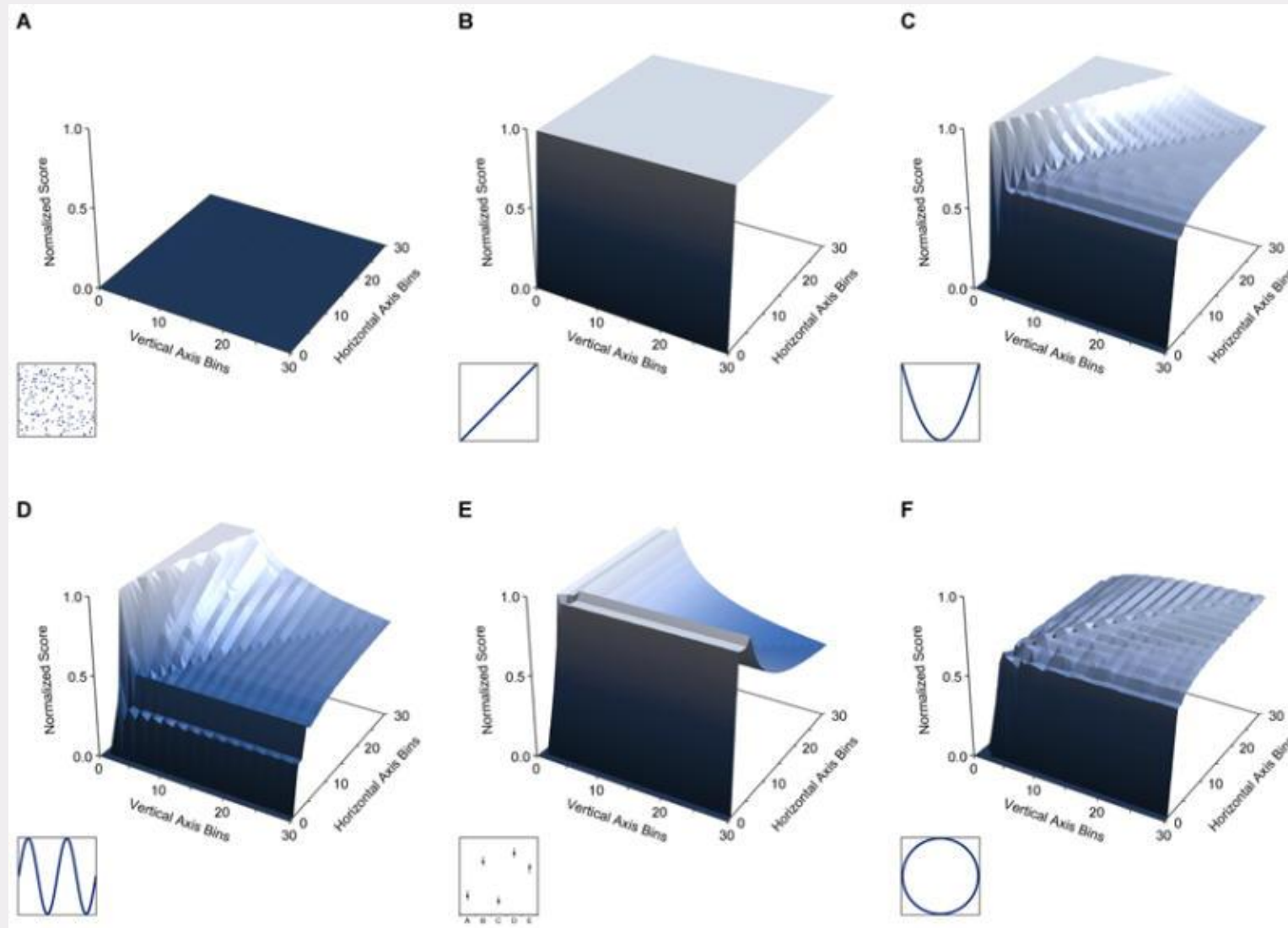
Metrics

Pearson

Problem

Contents

3. MIC



Metrics

Pearson

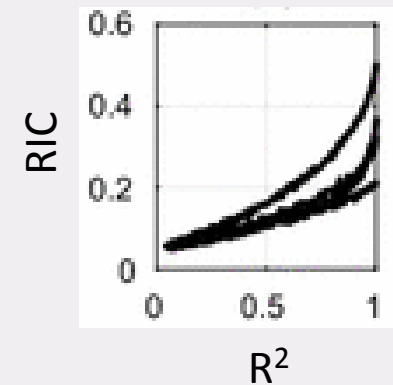
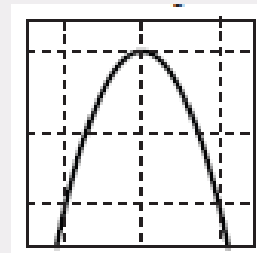
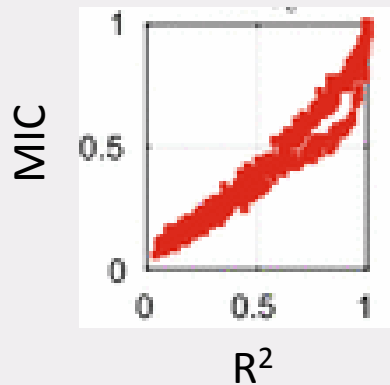
Problem

Contents

4. RIC

El Randomized Information Coefficient (RIC) es una medida similar a MIC, pero genera particiones aleatorias del rango de los datos en la variable X y en la variable Y.

Al igual que MIC, RIC es una métrica normalizada, donde $\text{RIC}=1$ es el máximo posible. Un valor cercano a 0 representa una débil o inexistente relación funcional entre las dos variables.



5. DISTANCE CORRELATION

Distance Correlation (dCor o dCorr) es una medida de dependencia entre dos vectores aleatorios X e Y . Los vectores aleatorios son independientes si y solo si el coeficiente dCorr es cero. dCorr pertenece a la misma familia de medidas de dependencia que PCC (basada en correlación).

$$a_{j,k} = \|X_j - X_k\|$$

$$b_{j,k} = \|Y_j - Y_k\|$$

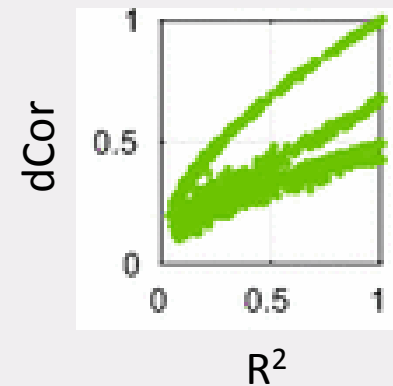
$$A_{j,k} := a_{j,k} - \bar{a}_{j.} - \bar{a}_{.k} + \bar{a}_{..}$$

$$B_{j,k} := b_{j,k} - \bar{b}_{j.} - \bar{b}_{.k} + \bar{b}_{..}$$

$$dCov^2(X, Y) = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n A_{j,k} B_{j,k}$$

$$dVar^2(X) := dCov^2(X, X) = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n A_{j,k}^2$$

$$dCor(X, Y) = \frac{dCov(X, Y)}{\sqrt{dVar(X) \cdot dVar(Y)}}$$



6. BIWEIGHT MIDCORRELATION

Biweight Midcorrelation (BiCor) es una medida de similitud entre muestras. BiCor está basado en la mediana, a diferencia de PCC, lo cual la hace menos sensible a *outliers*.

$$u_i = \frac{x_i - \text{med}(x)}{9 \cdot \text{mad}(x)} \quad v_i = \frac{y_i - \text{med}(y)}{9 \cdot \text{mad}(y)}$$

$$w_{i,x} = (1 - u_i^2)^2 \cdot I(1 - |u_i|)$$

$$w_{i,y} = (1 - v_i^2)^2 \cdot I(1 - |v_i|)$$

$$I(x) = \begin{cases} 1, & x > 0 \\ 0, & \text{de otro modo} \end{cases}$$

$$\bar{x}_i = \frac{(x_i - \text{med}(x)) \cdot w_{i,x}}{\sqrt{\sum_{j=1}^n [(x_j - \text{med}(x)) \cdot w_{j,x}]^2}}$$

$$\bar{y}_i = \frac{(y_i - \text{med}(y)) \cdot w_{i,y}}{\sqrt{\sum_{j=1}^n [(y_j - \text{med}(y)) \cdot w_{j,y}]^2}}$$

$$BICOR(x, y) = \sum_{i=1}^n \bar{x}_i \bar{y}_i$$



7. METODOLOGÍA MUTUAL RANK

Obayashi y Kinoshita muestran que es mejor utilizar el ranking de los valores de correlación que utilizar los valores de correlación en las redes de coexpresión.

DNA RESEARCH **16**, 249–260, (2009)

doi:10.1093/dnares/dsp016

Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression

TAKESHI Obayashi¹ and KENGO Kinoshita^{1,2,*}

Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan¹ and Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan²

(Received 30 March 2009; accepted 14 August 2009; published online 18 September 2009)

Abstract

Information regarding gene coexpression is useful to predict gene function. Several databases have been constructed for gene coexpression in model organisms based on a large amount of publicly available gene expression data measured by GeneChip platforms. In these databases, Pearson's correlation coefficients (PCCs) of gene expression patterns are widely used as a measure of gene coexpression. Although the coexpression measure or GeneChip summarization method affects the performance of the gene coexpression database, previous studies for these calculation procedures were tested with only a small number of samples and a particular species. To evaluate the effectiveness of coexpression measures, assessments with large-scale microarray data are required. We first examined characteristics of PCC and found that the optimal PCC threshold to retrieve functionally related genes was affected by the method of gene expression database construction and the target gene function. In addition, we found that this problem could be overcome when we used correlation ranks instead of correlation values. This observation was evaluated by large-scale gene expression data for four species: *Arabidopsis*, human, mouse and rat.



Ranking

Metrics

Pearson

Problem

Contents

Results

7. METODOLOGÍA MUTUAL RANK

Ranking for Gene A		
Rank	Gene	PCC
0	A	1,000
1	B	0,995
2	E	0,985
3	R	0,982
4	S	0,980
5	C	0,971

Ranking for Gene B		
Rank	Gene	PCC
0	B	1,000
1	T	0,997
2	A	0,993
3	R	0,989
4	S	0,982
5	C	0,980

Ranking for Gene C		
Rank	Gene	PCC
0	C	1,000
1	J	0,991
2	B	0,980
3	K	0,977
4	A	0,971
5	U	0,968

$$R_{A \rightarrow A} = R_{B \rightarrow B} = R_{C \rightarrow C} = 0$$

$$R_{A \rightarrow B} = 1 \quad R_{A \rightarrow C} = 5$$

$$R_{B \rightarrow A} = 2 \quad R_{B \rightarrow C} = 5$$

$$R_{C \rightarrow A} = 4 \quad R_{C \rightarrow B} = 2$$

$$MR_{A,B} = \sqrt{R_{A \rightarrow B} \cdot R_{B \rightarrow A}} = \sqrt{1 \cdot 2} = 1,414$$

$$MR_{A,C} = \sqrt{R_{A \rightarrow C} \cdot R_{C \rightarrow A}} = \sqrt{5 \cdot 4} = 4,472$$

$$MR_{B,C} = \sqrt{R_{B \rightarrow C} \cdot R_{C \rightarrow B}} = \sqrt{5 \cdot 2} = 3,162$$



Ranking

Metrics

Pearson

Problem

Contents

Results

7. METODOLOGÍA MUTUAL RANK

A partir de los valores de Mutual Rank (MR), se procede a generar la red de coexpresión genética definiendo un umbral por encima del cual no se considera significativa la relación en los perfiles de expresión de dos genes.

La literatura sugiere diferentes valores para este umbral, tales como $MR=20$, $MR=100$ o MR tal que sólo el 1% de arcos sean escogidos.

En este estudio se escogió **MR=20** como umbral (Obayashi & Kinoshita, 2010)



Ranking

Metrics

Pearson

Problem

Contents

8. RESULTADOS Y DISCUSIÓN

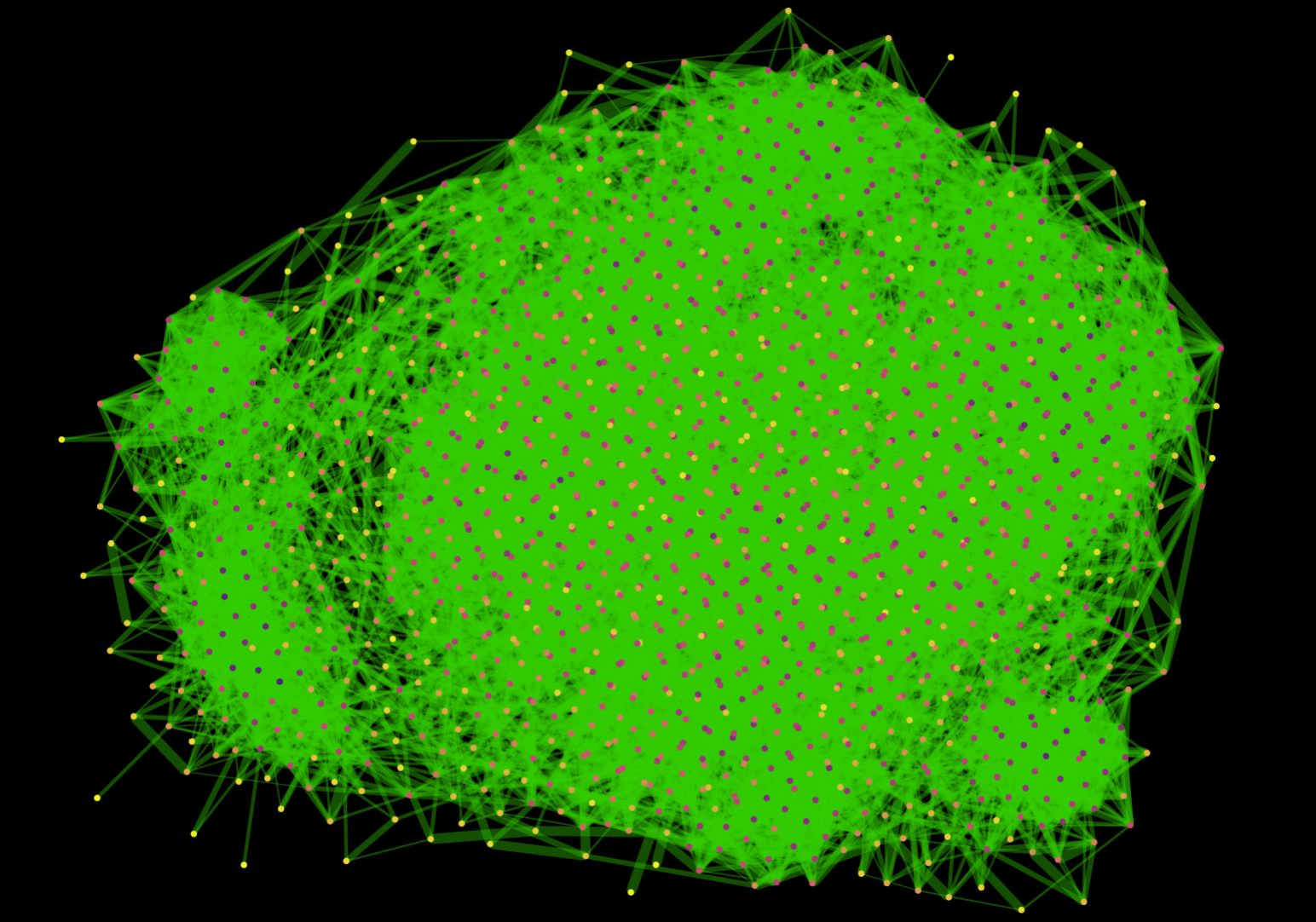
Al realizar la implementación de las métricas y el cálculo de las mismas para los 1600 genes aleatoriamente escogidos, se generaron las respectivas redes de coexpresión encontraron las siguientes medidas:

Métrica	Arcos	Nodos sin conexión	Comunidades	Componentes	Nodos en componente más grande
PCC	12 489	0	8	1	1600
MIC	7 755	53	22	63	1528
RIC	8 097	60	20	69	1522
dCor	8 263	9	7	10	1600
BICOR	2 336	935	8	937	585



8. RESULTADOS Y DISCUSIÓN

RED DE COEXPRESIÓN CON PEARSON



Results

Ranking

Metrics

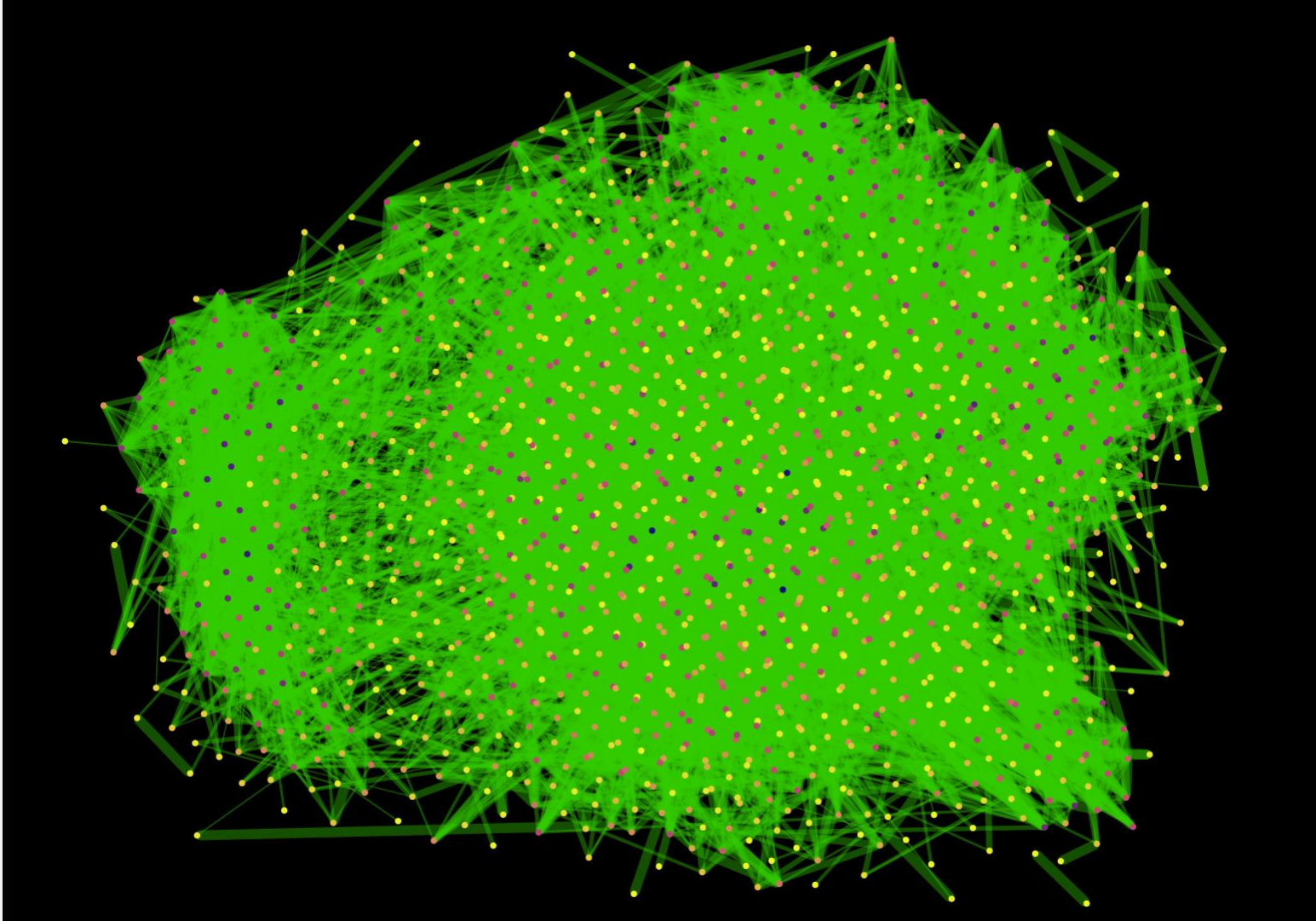
Pearson

Problem

Contents

8. RESULTADOS Y DISCUSIÓN

RED DE COEXPRESIÓN CON MAXIMAL INFORMATION COEFFICIENT



Results

Ranking

Metrics

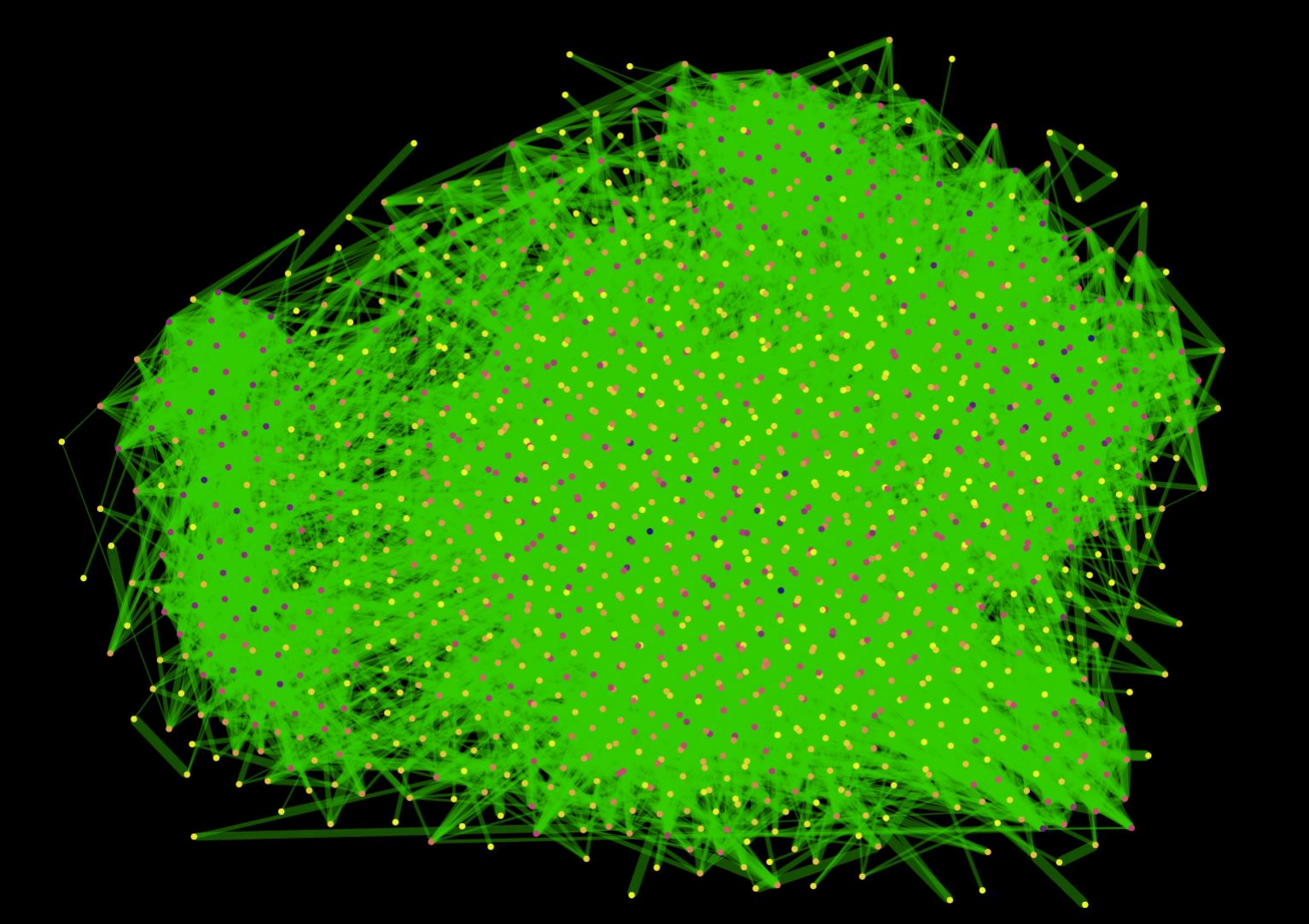
Pearson

Problem

Contents

8. RESULTADOS Y DISCUSIÓN

RED CON RANDOMIZED INFORMATION COEFFICIENT



Results

Ranking

Metrics

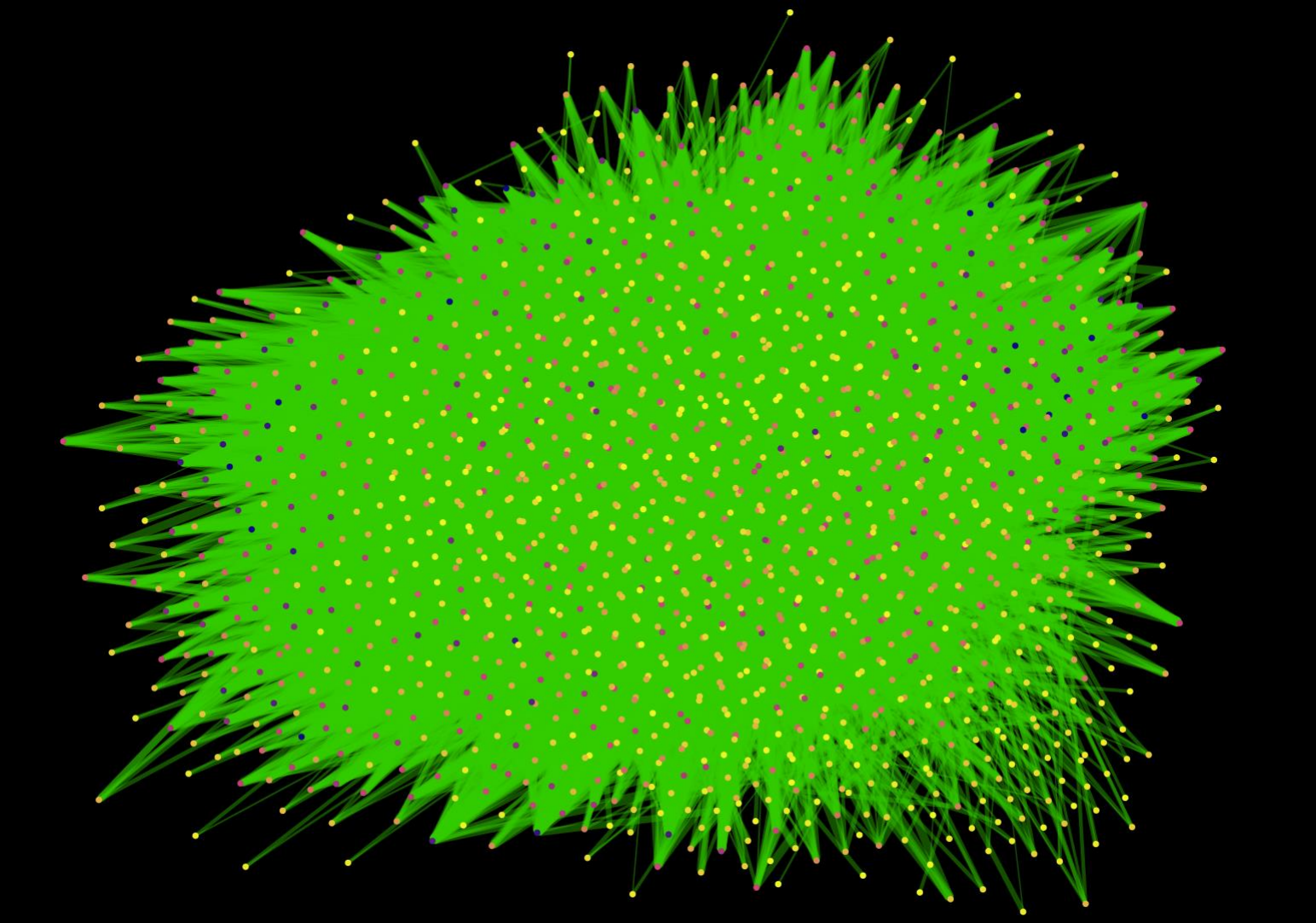
Pearson

Problem

Contents

8. RESULTADOS Y DISCUSIÓN

RED DE COEXPRESIÓN CON DISTANCE CORRELATION



Results

Ranking

Metrics

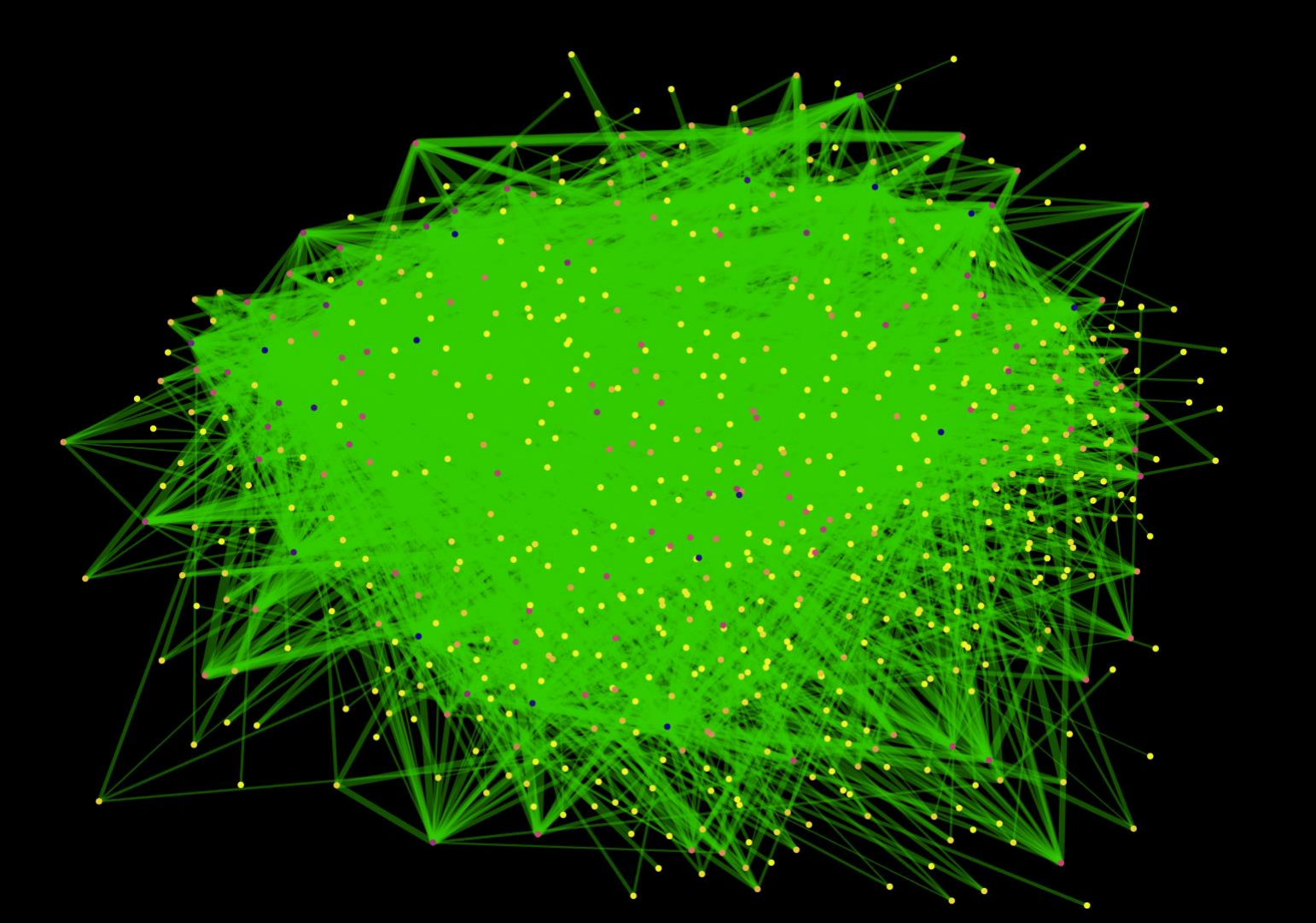
Pearson

Problem

Contents

8. RESULTADOS Y DISCUSIÓN

RED DE COEXPRESIÓN CON BIWEIGHT MIDCORRELATION



Results

Ranking

Metrics

Pearson

Problem

Contents

8. RESULTADOS Y DISCUSIÓN

A continuación un resumen de la complejidad computacional del cálculo de cada métrica y el tiempo aproximado que tardó para 1600 genes:

• PCC:	$O(n)$	<10min
• MIC:	$O(n^{3.6})$	~15h
• RIC*:	$O(n^{1.5})$	~8h
• dCorr:	$O(n^2)$	~3h
• BICOR**:	$O(n)$	<10min



Results

Ranking

Metrics

Pearson

Problem

Contents

MUCHAS GRACIAS

**No act of kindness, no matter
how small, is ever wasted.**

Aesop

