CICESE | Maestría en Ciencias de la Computación | Reconocimiento de Patrones

Biclustering para datos de Expresión Genética utilizando Optimización Multiobjetivo MOEA/D y posterior análisis mediante SOM y PCA.

Estudiantes:

- César Miguel Valdez Córdova
- Luis Enrique García Hernández

Indice

- El Problema de Biclustering
- MOEA/D
- Análisis de Componentes Principales (PCA)
- Mapa auto-organizado (SOM)
- Resultados
- Conclusiones y Trabajo a Futuro

Objetivos

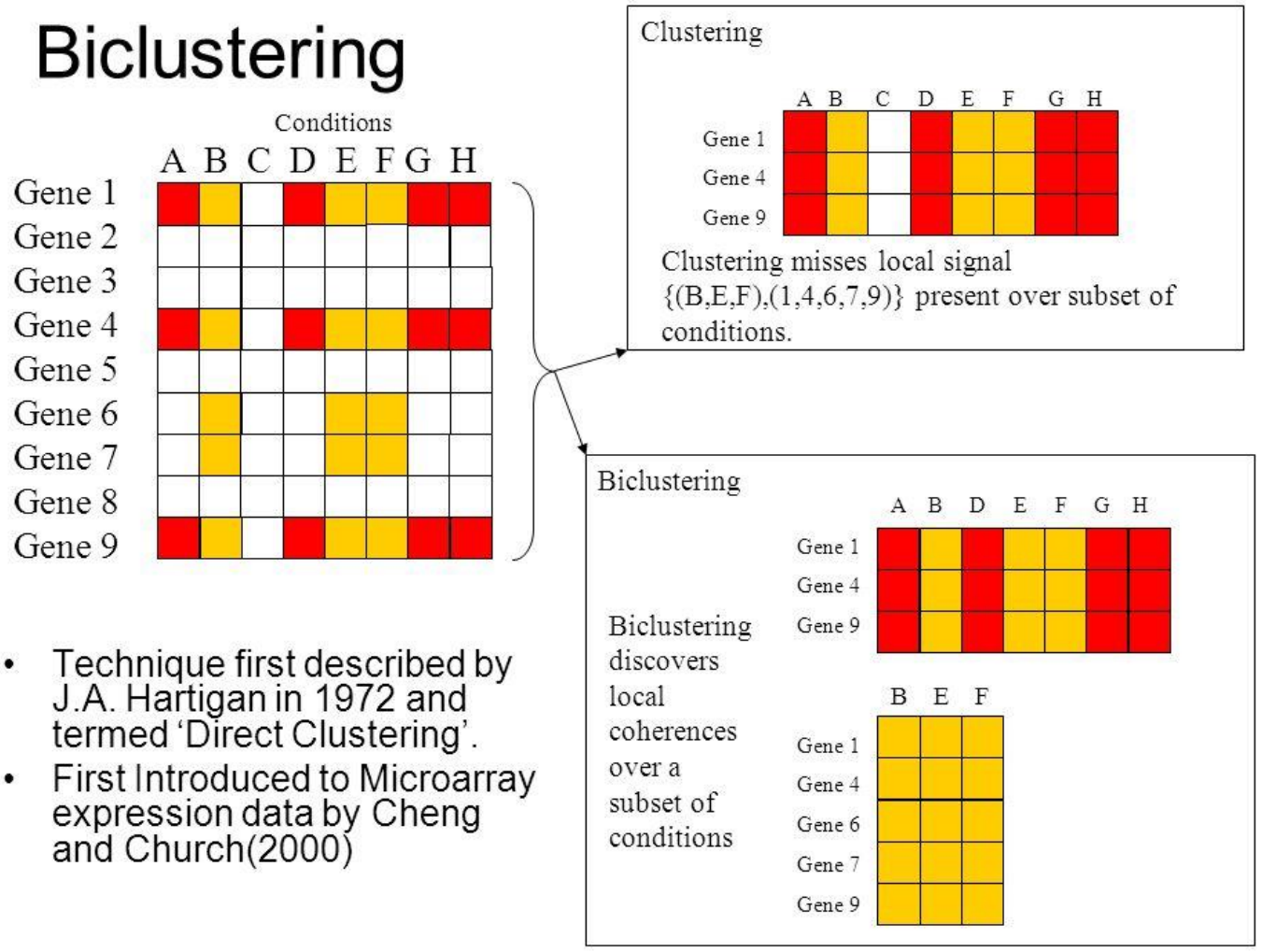
Objetivo General: Utilizar el algoritmo genético multi-objetivo, MOEA/D, para la identificación de biclusters de interés en datos de Expresión Genética y posterior análisis mediante SOM y PCA.

Objetivos específicos:

- Implementar el algoritmo genético multi-objetivo: MOEA/D para la identificación de biclusters.

- Realizar análisis mediante las técnicas de análisis antes mencionadas y determinar si las reducciones encontradas por dichos métodos están relacionadas con los biclusters generados.

Biclustering



- Technique first described by J.A. Hartigan in 1972 and termed 'Direct Clustering'.
- First Introduced to Microarray expression data by Cheng and Church(2000)

Definición del problema

- ¿Por qué usar el biclustering para resolver este problema?
- Permite detectar grupos traslapados entre los biclusters, proporcionando una mejor representación de la realidad biológica que implica genes con muchas funciones o reguladas por muchos factores.

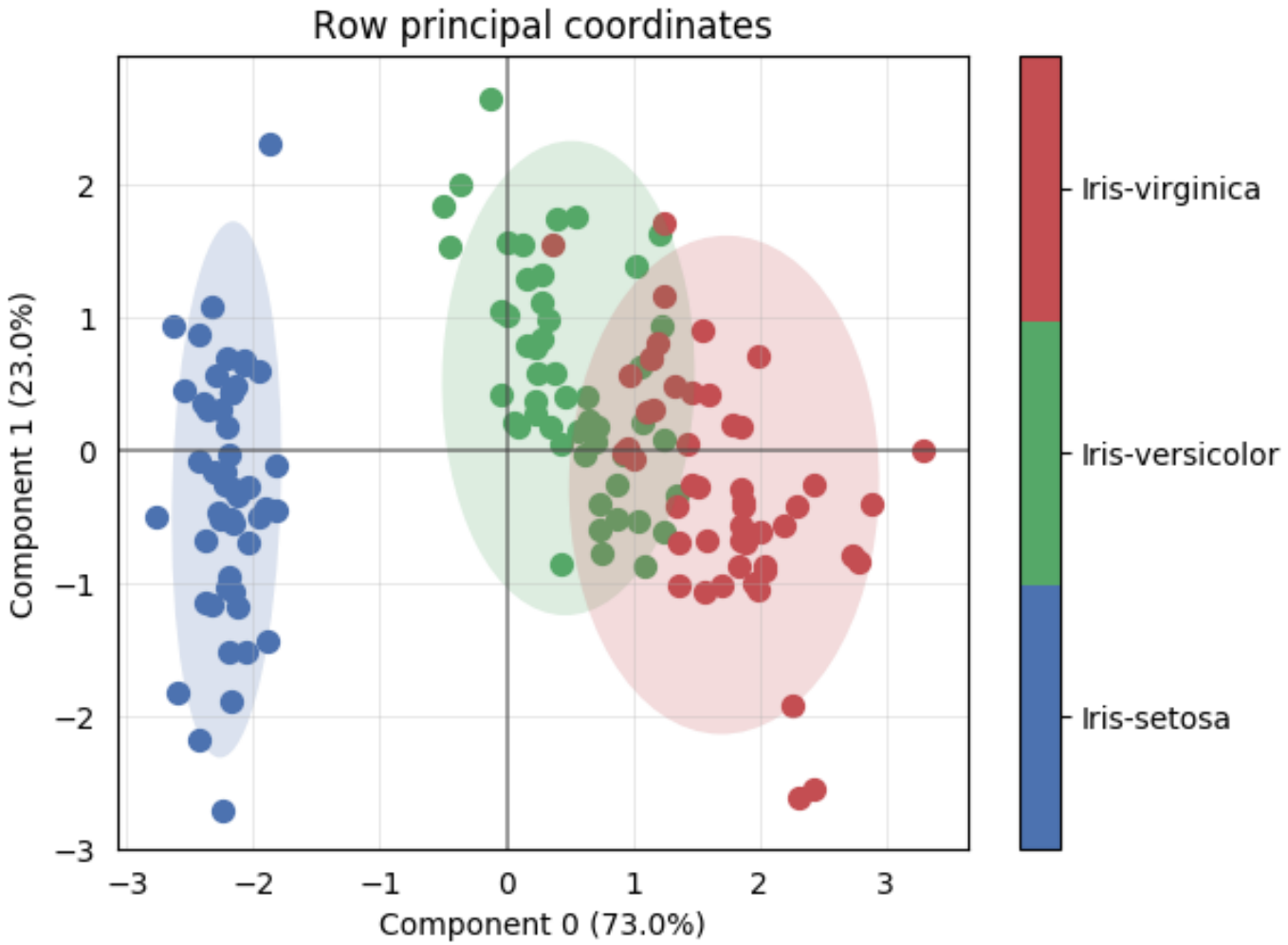
Complejidad del problema

- El biclustering es un problema NP-difícil (Demostrado en el 2002 Tanay et al.), y es por ello que la mayoría de los métodos utilizados se basan en procedimientos de optimización. El que sea NP-difícil implica que una búsqueda exhaustiva en el espacio de decisión no sea factible, pero aplicando una medida de calidad a una solución candidata, el uso de una meta-heurística para la solución del problema parece apropiado.

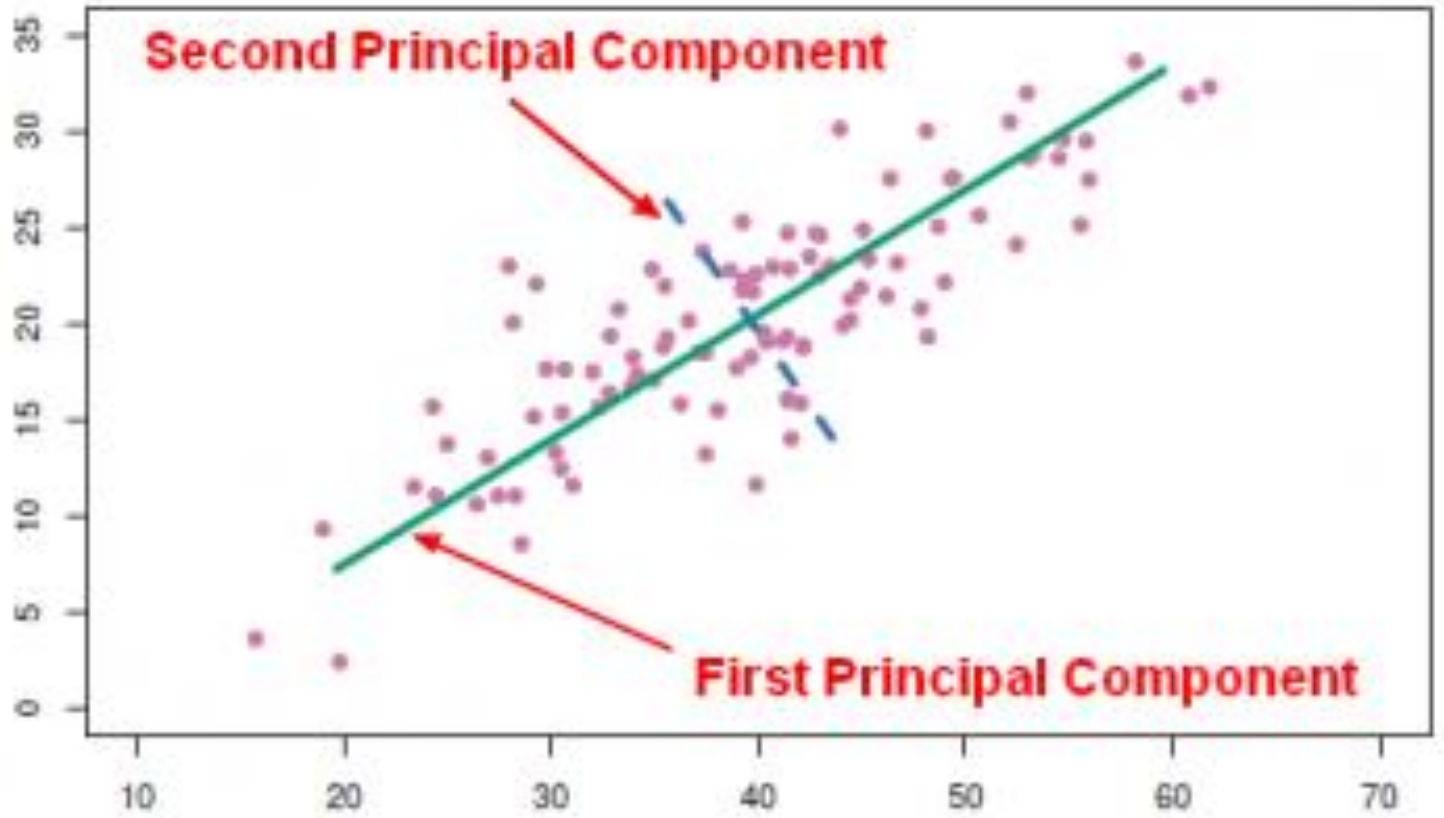
Análisis de componentes principales (PCA)

- El análisis de componentes principales (PCA) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos.
- PCA busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Esta convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales.

Análisis de componentes principales (PCA)

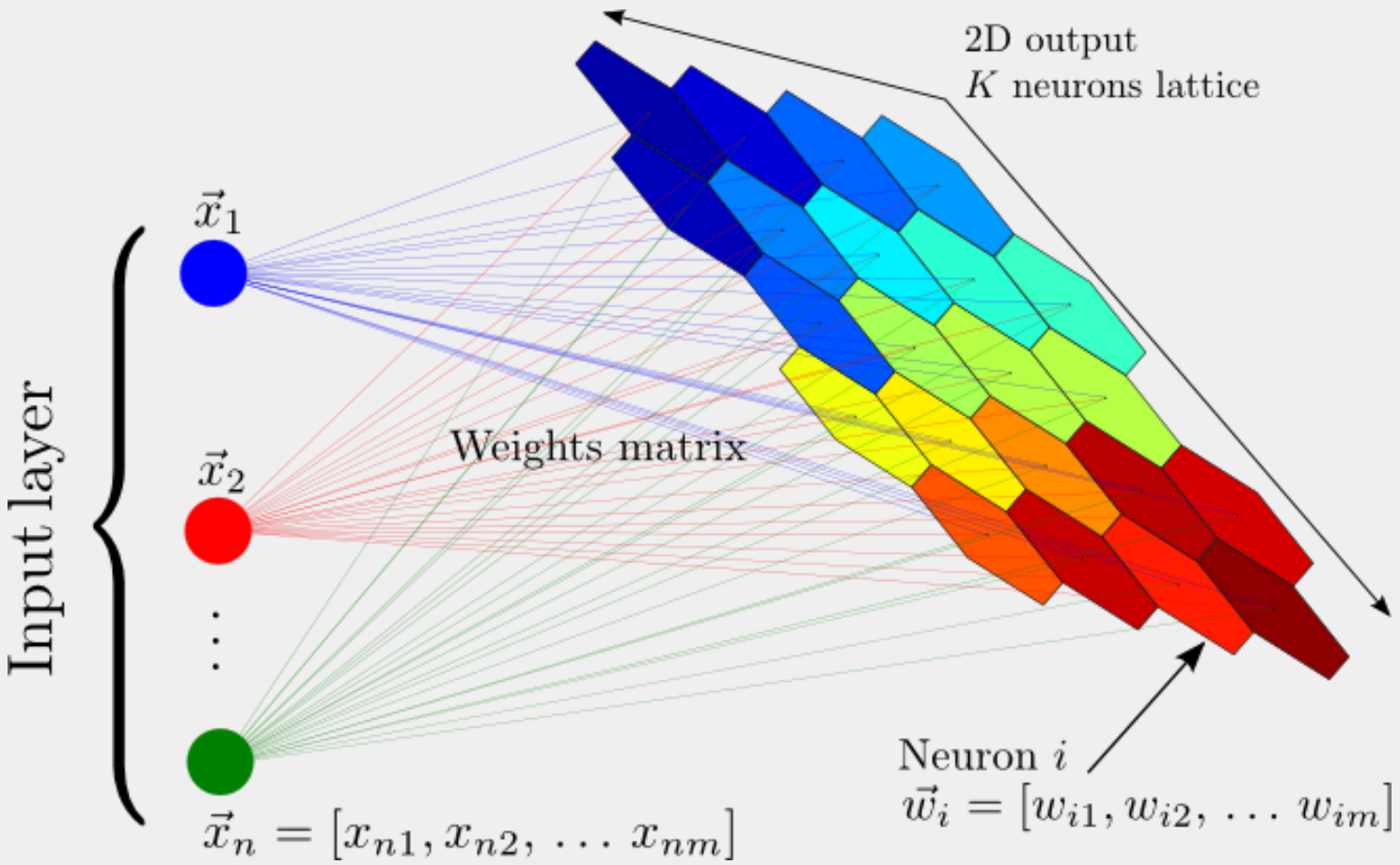


Análisis de componentes principales (PCA)

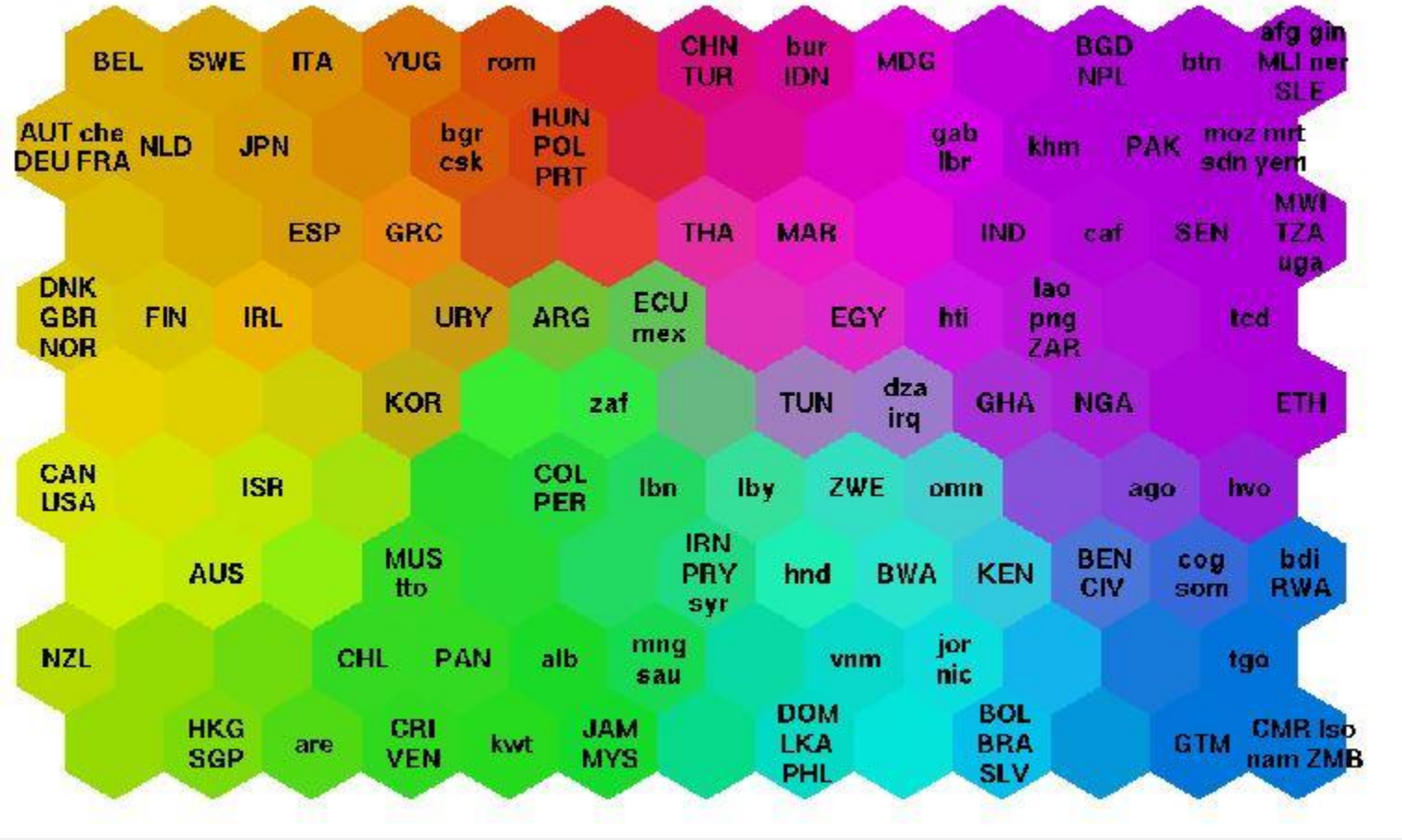


Mapas auto-organizados (SOM)

- Un mapa auto-organizado (SOM) es un tipo de red neuronal artificial, que es entrenada usando aprendizaje no supervisado para producir una representación discreta del espacio de las muestras de entrada, llamado mapa. Los mapas auto-organizados usan una función de vecindad para preservar las propiedades topológicas del espacio de entrada.
- Los SOMs son útiles para visualizar vistas de baja dimensión de datos de alta dimensión


2D output
K neurons lattice
$\vec{x}_1$
$\vec{x}_2$
Input layer
Weights matrix
$\vec{x}_n = [x_{n1}, x_{n2}, \ldots x_{nm}]$
Neuron i
$\vec{w}_i = [w_{i1}, w_{i2}, \ldots w_{im}]$

Countries organized on a self-organizing map based on indicators related to poverty:



Datos Utilizados

Conjuntos de datos de expresión genética de referencia

- Levadura Saccharomyces Cerevisiae (con 2884 genes, 17 condiciones)
- Linfoma de células B humanas (4026 genes y 96 condiciones)
- Datos de expresión de genes de 14 tipos de cáncer. (propuesto por Ramaswamy et. al.[7])

Referencias:

http://arep.med.harvard.edu/biclustering/

https://github.com/probml/pmtk3/tree/master/bigData/14cancer

MOEAD/D|

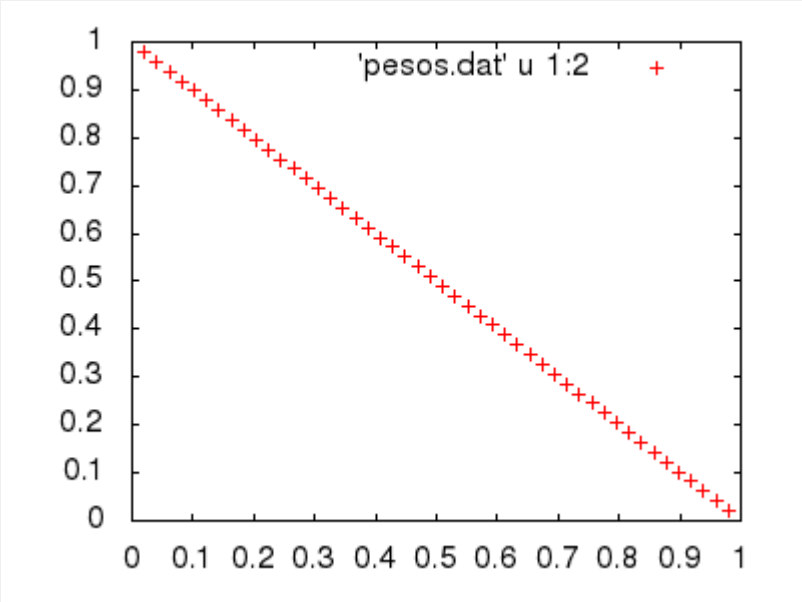
Implementación del algoritmo

MOEA/D

- Algoritmo evolutivo multiobjetivo
- Basado en descomposición
- Entradas
  - Funciones objetivo a evaluar
  - Criterio de parada
  - N: Numero de subproblemas
  - Dispersión uniforme de vectores de peso
  - T: Tamaño del vecindario
- Salida
  - Frente Pareto

Algoritmo implementado

1. Inicialización
    1. Calcular la distancia Euclidiana entre dos vectores de pesos y luego calcular los vectores más cercanos a cada vector de peso.



$$\sum_{i=1}^{n} \lambda_i = 1$$

Algoritmo implementado

1.3. Generar una población inicial $x^1, \ldots, x^N$ de manera aleatoria.

	1	2	3	4	5
1	130	20	65	31	4
2	38	42	15	16	70
3	20	61	110	83	96
4	92	75	16	105	85
5	48	77	90	113	41
6	17	8	100	37	78
7	101	40	65	33	15
8	28	13	117	38	19
9	121	28	97	43	50
10	72	86	34	97	42
11	24	14	7	102	17
12	93	100	45	30	3
13	48	36	83	63	79
14	86	117	2	55	36
15	95	48	68	41	15

Matriz de expresión de genes

G	3	5	6	8	10	12
C	2	3	5			

Representacion

	2	3	5
3	61	110	96
5	77	90	41
6	8	100	78
8	13	117	19
12	100	45	3

Bicluster

Algoritmo implementado

2. Actualización

   For i = 1 ,..., N

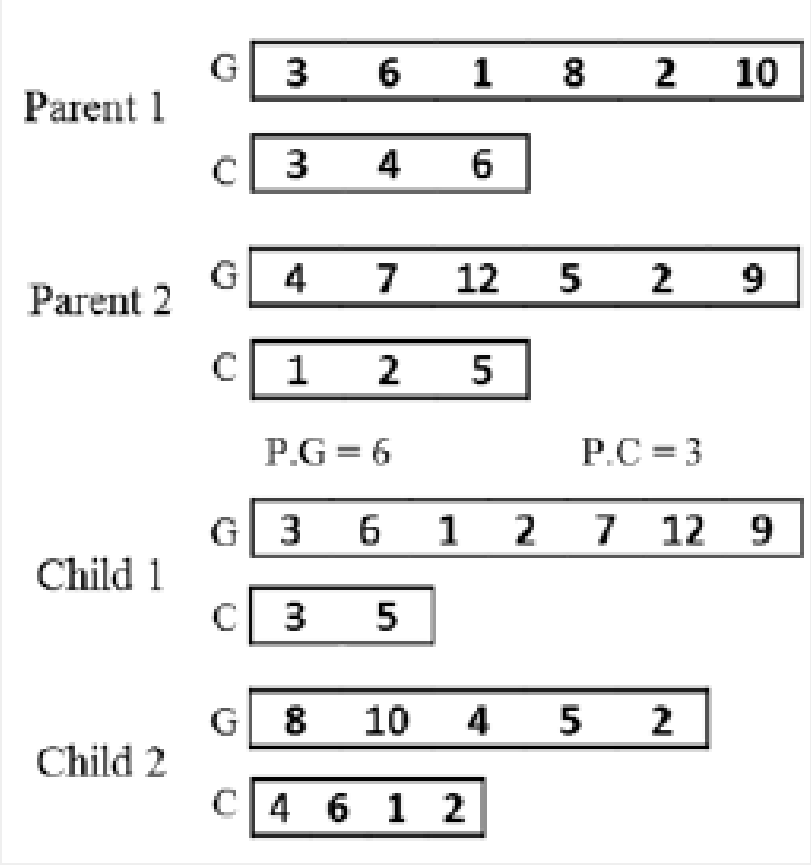
      2.1 Selección de padres

      2.2 Cruzamiento

      2.3 Mutación

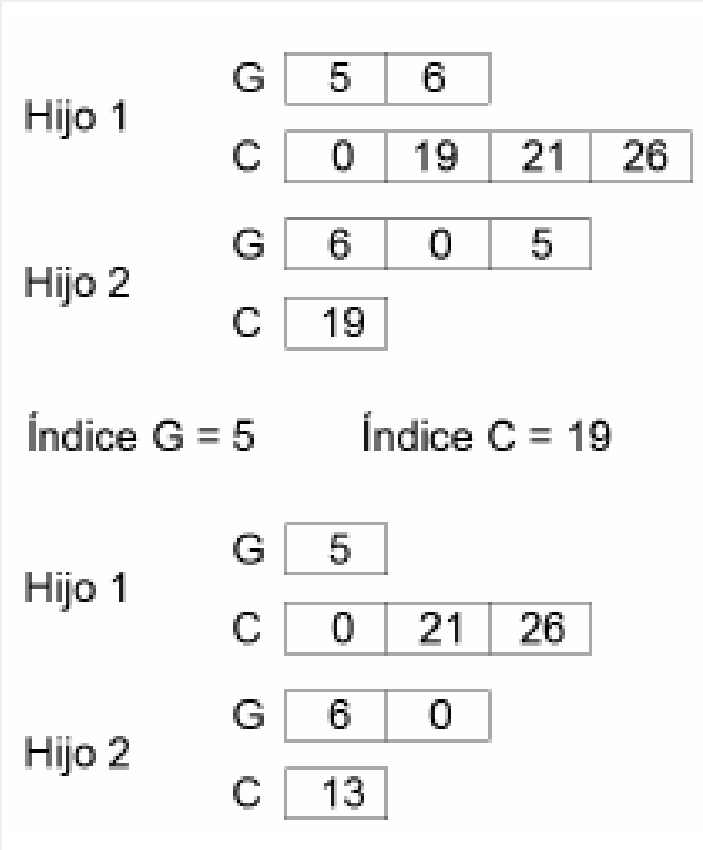
Algoritmo implementado

2.2. Cruzamiento



Algoritmo implementado

2.3. Mutación



Algoritmo implementado

2.4. Mejora utilizando conocimiento del problema

2.5. Actualización del vecindario.

$$g^{te}(x\ \lambda^j, z^*) = max_{1 \leq i \leq m}\{\lambda_i^j | f_i(x) - z_{i|}^*\}$$

x = Vector solución
λ = Vector de pesos
m = Número de funciones objetivo
z^* = Vector de referencia

Para cada índice , si ent.

$j \in B(i)$ $g^{te}(y'|\lambda^i, z) \leq g^{te}(x^j|\lambda^j, z)$

$x^j = y'$

Algoritmo implementado

2.7. Actualizar EP

- El MSR de la solución i es menor o igual que el MSR de la solución j, y el tamaño de la solución i es mayor que la solución j.
- El tamaño de la solución i es mayor o igual al tamaño de la solución j, y el MSR de la solución i es menor que el de la solución j.

➔ Se remueven del frente aquellos vectores dominados por y’

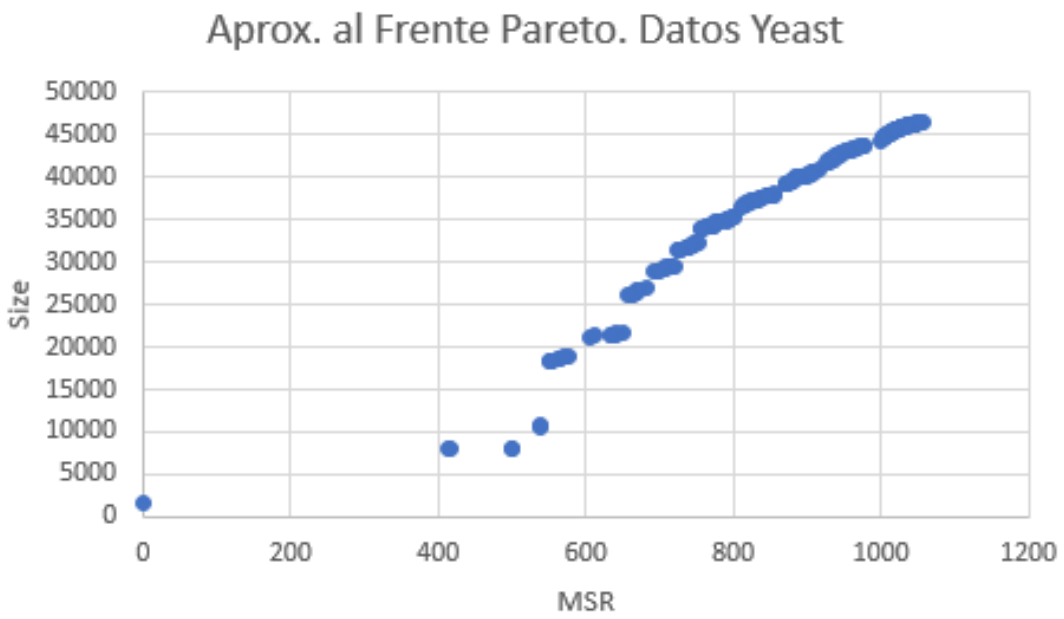
➔ Se agrega y+ al frente si no existen vectores en el frente que lo dominen.

Funciones objetivo MOEA/D

- Se utilizarán 2 funciones objetivo, en donde buscamos maximizar el tamaño del bicluster y minimizar el MSR (Mean Squared Residue). El MSR es una medida que evalúa la coherencia entre genes y condiciones.
- Estas dos funciones se encuentran en conflicto ya que entre más grande sea el bicluster, es más probable que aumente el valor de MSR.

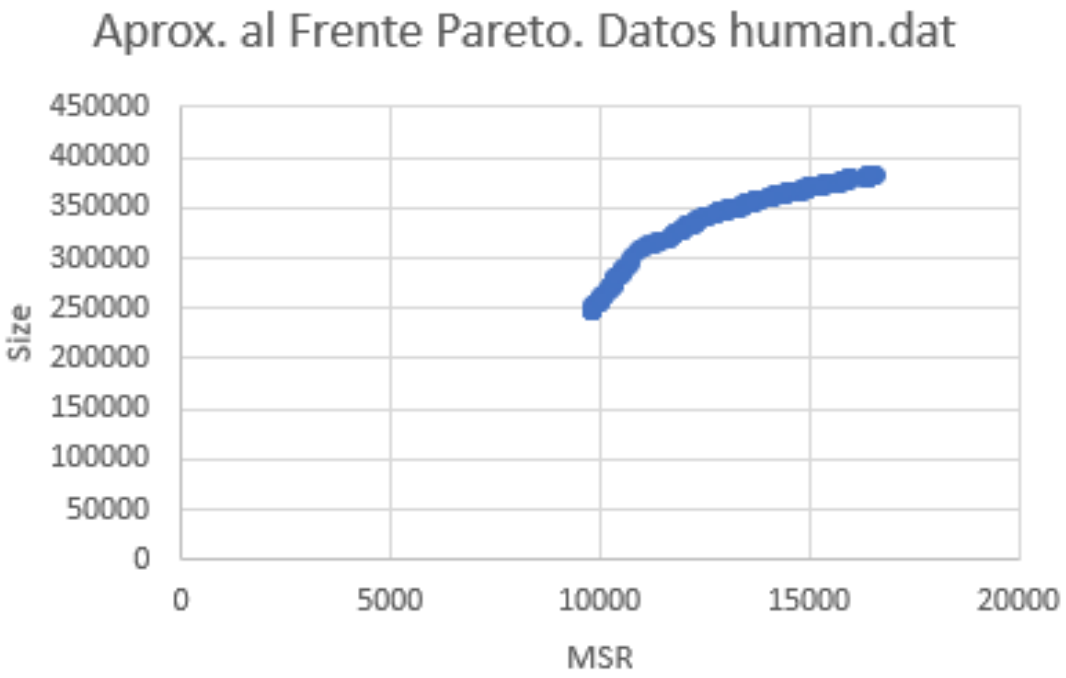
Experimentos y resultados

Resultados datos de levadura



Promedio	Resultados	Referencia
Tiempo de ejecución	60 min	70 min
MSR	853	832
Tamaño	39956	40236

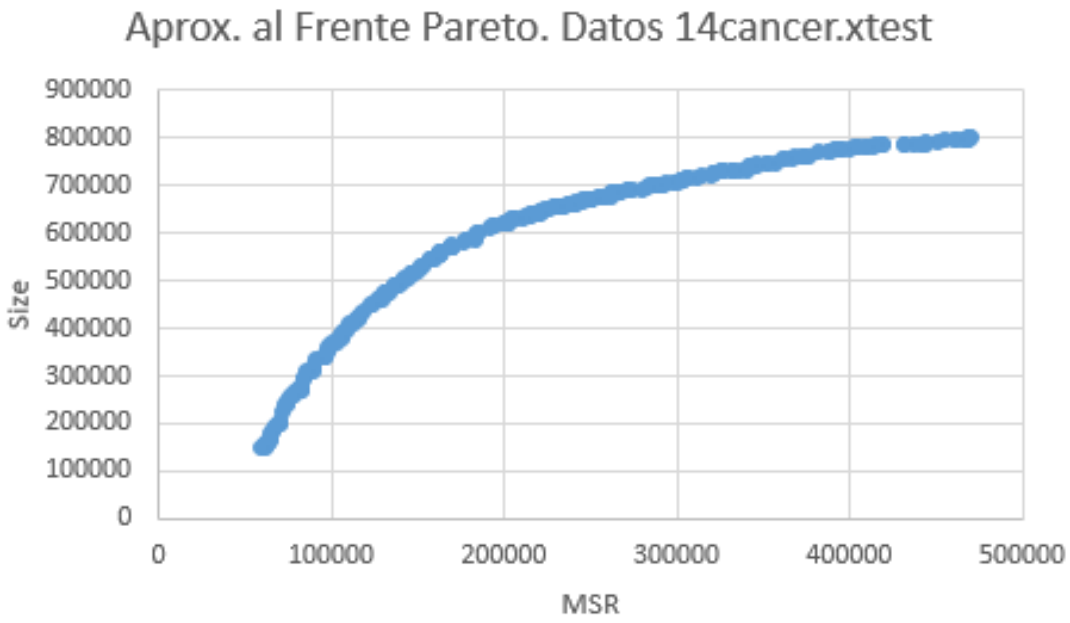
Resultados datos Linfoma de células B humanas



Promedio	Resultados	Referencia
Tiempo de ejecución	75 min	70 min
MSR	13242	13168
Tamaño	338051	349654

Elemento, O., Slonim, N., & Tavazoie, S. (2007). A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Molecular Cell, 28(2), 337–350. https://doi.org/10.1016/j.molcel.2007.09.027

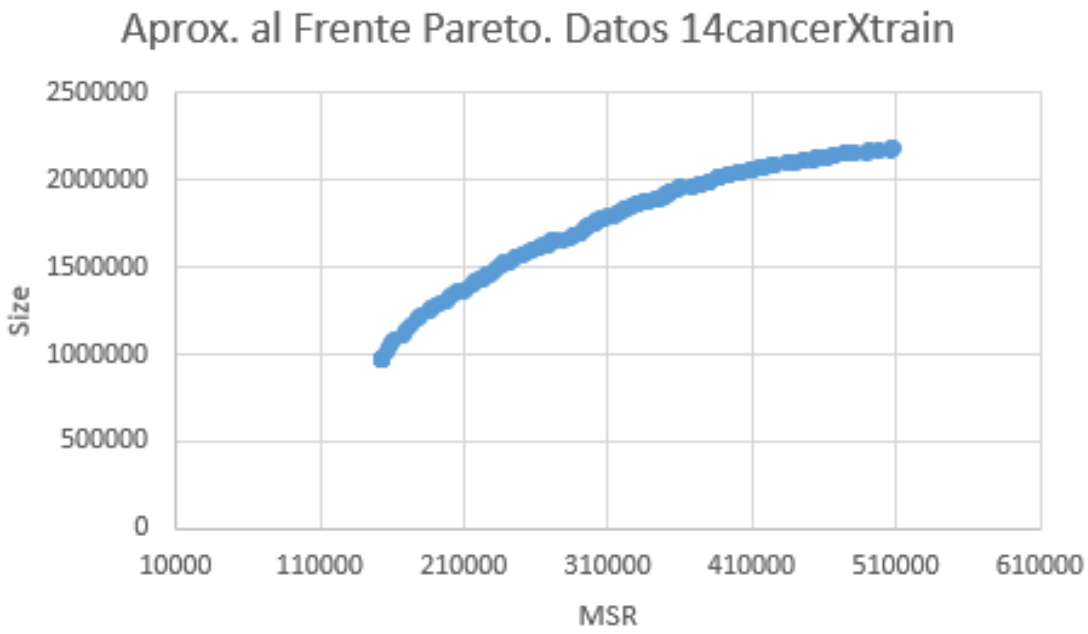
Resultados datos 14cancer.xtest



Promedio	Resultados	Referencia
Tiempo de ejecución	80 min	75 min
MSR	225780	228980
Tamaño	575659	582398

Elemento, O., Slonim, N., & Tavazoie, S. (2007). A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Molecular Cell, 28(2), 337–350. https://doi.org/10.1016/j.molcel.2007.09.027

Resultados datos 14cancer.xtrain



Promedio	Resultados	Referencia
Tiempo de ejecución	120 min	115 min
MSR	319595	306987
Tamaño	1728704	1735289

Elemento, O., Slonim, N., & Tavazoie, S. (2007). A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Molecular Cell, 28(2), 337–350. https://doi.org/10.1016/j.molcel.2007.09.027

Parámetros

Tamaño de población	300
Generaciones	400
Tamaño del vecindario	30
Límite de soluciones actualizadas	5
ŏ	200
Probabilidad de cruza	1.0
Probabilidad de mutación	0.4
Probabilidad de mutación de genes	0.8
Probabilidad de mutación de condiciones	0.2

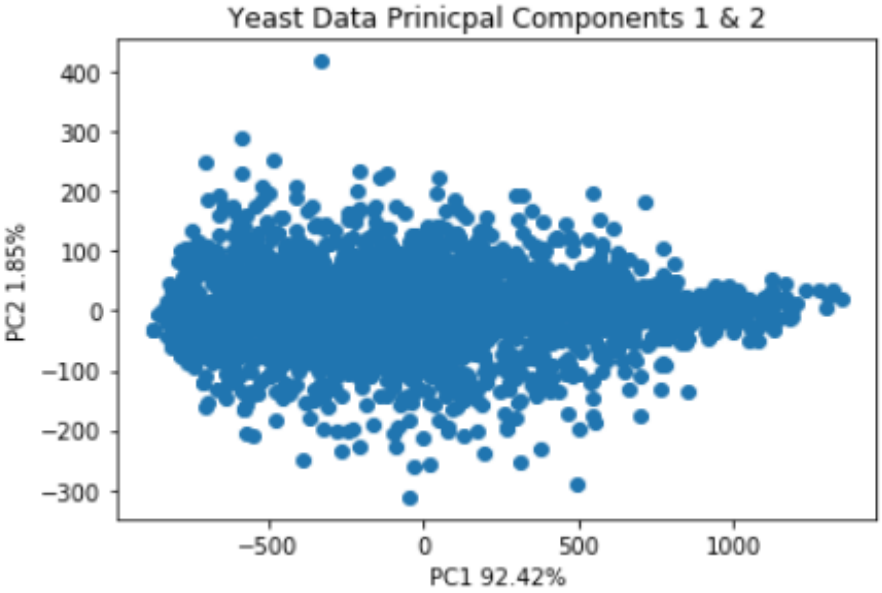
Resultados PCA

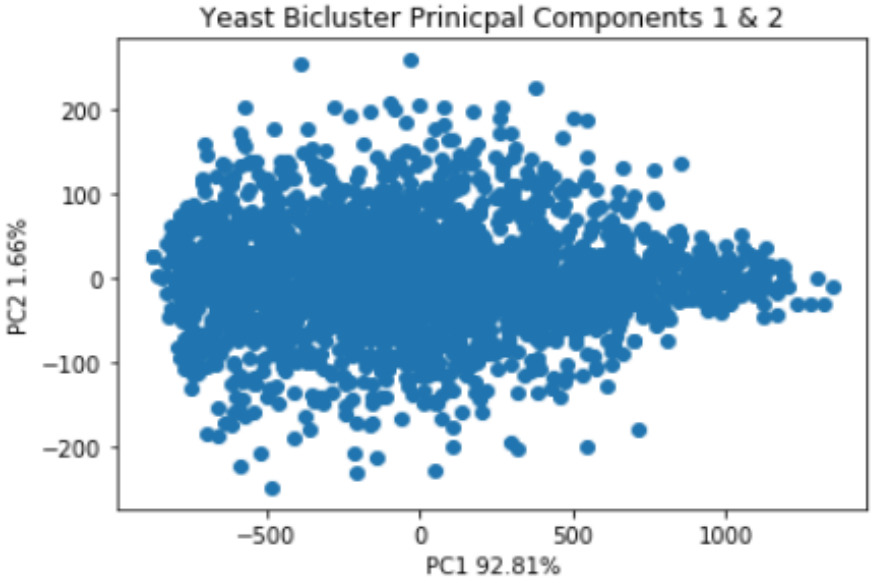
Varianza Explicada por Componentes Principales 1 & 2 en Datos Originales

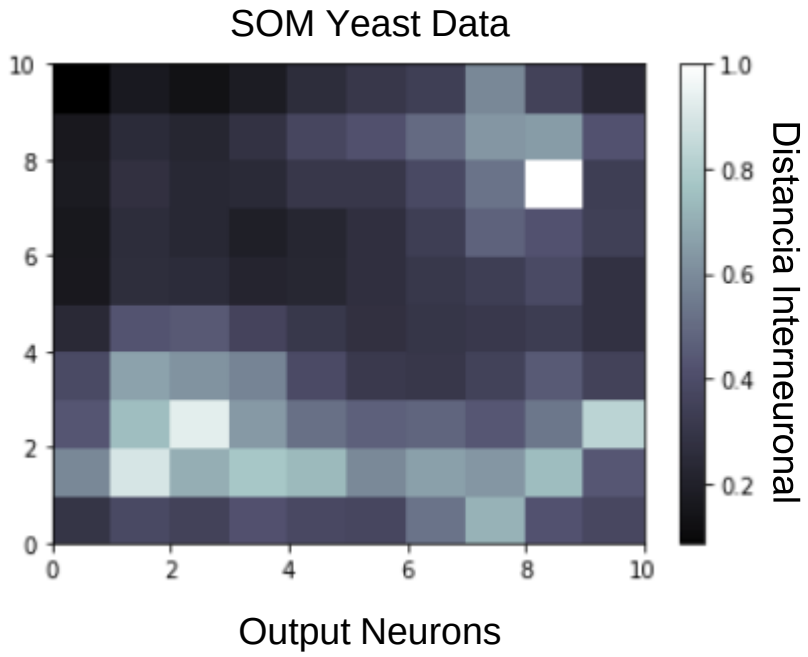
	PC1	PC2
Levadura	0.924	0.018
Humano	0.069	0.052
Cáncer	0.459	0.133

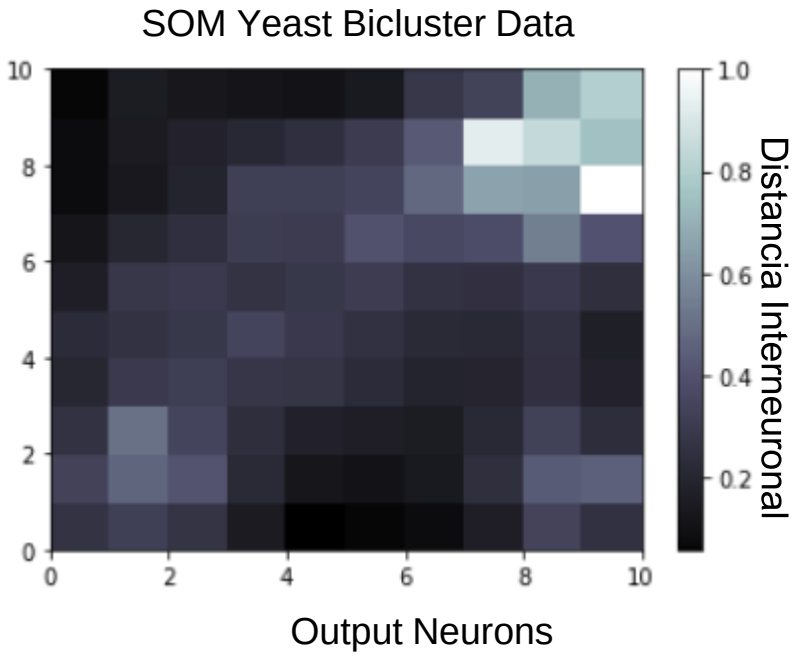
Varianza Explicada por Componentes Principales 1 & 2 en Biclusters

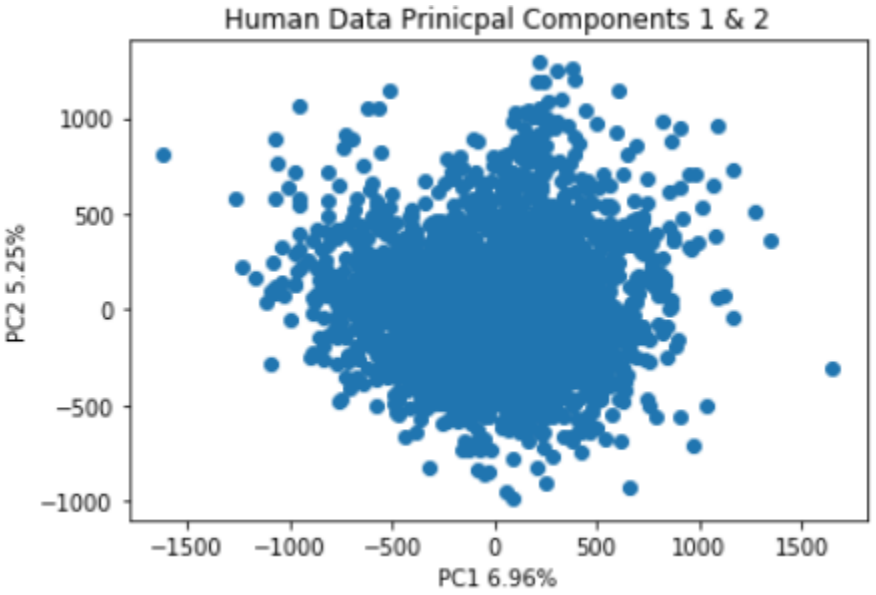
	PC1	PC2
Levadura	0.928	0.016
Humano	0.0714	0.050
Cáncer	0.437	0.122

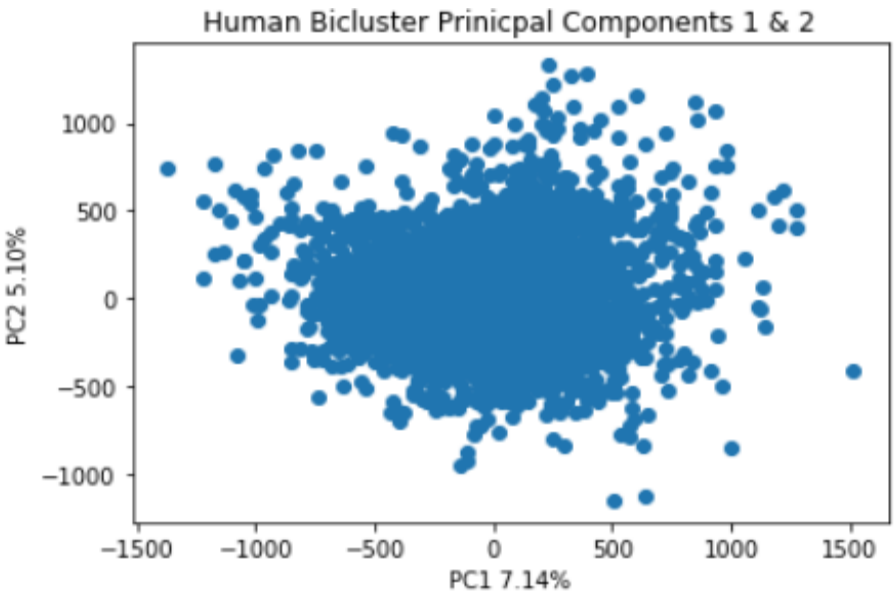

Yeast Data Prinicpal Components 1 & 2
PC2 1.85%
PC1 92.42%

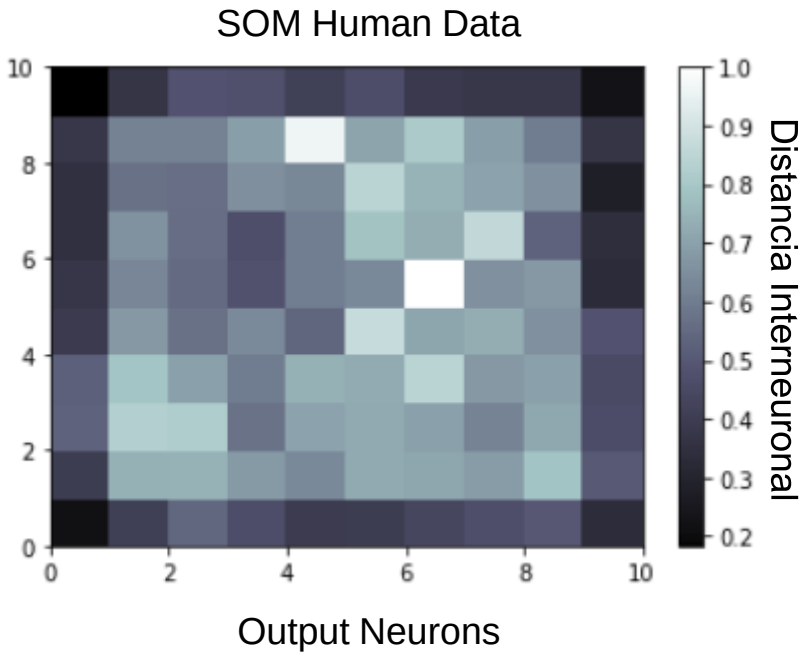

Yeast Bicluster Prinicpal Components 1 & 2
PC2 1.66%
PC1 92.81%

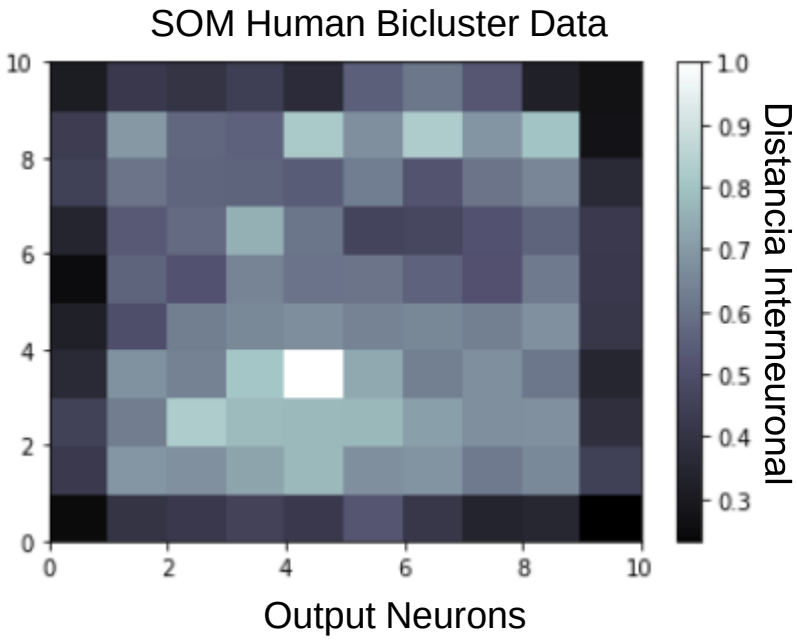

SOM Yeast Data
Distancia Interneuronal
Output Neurons

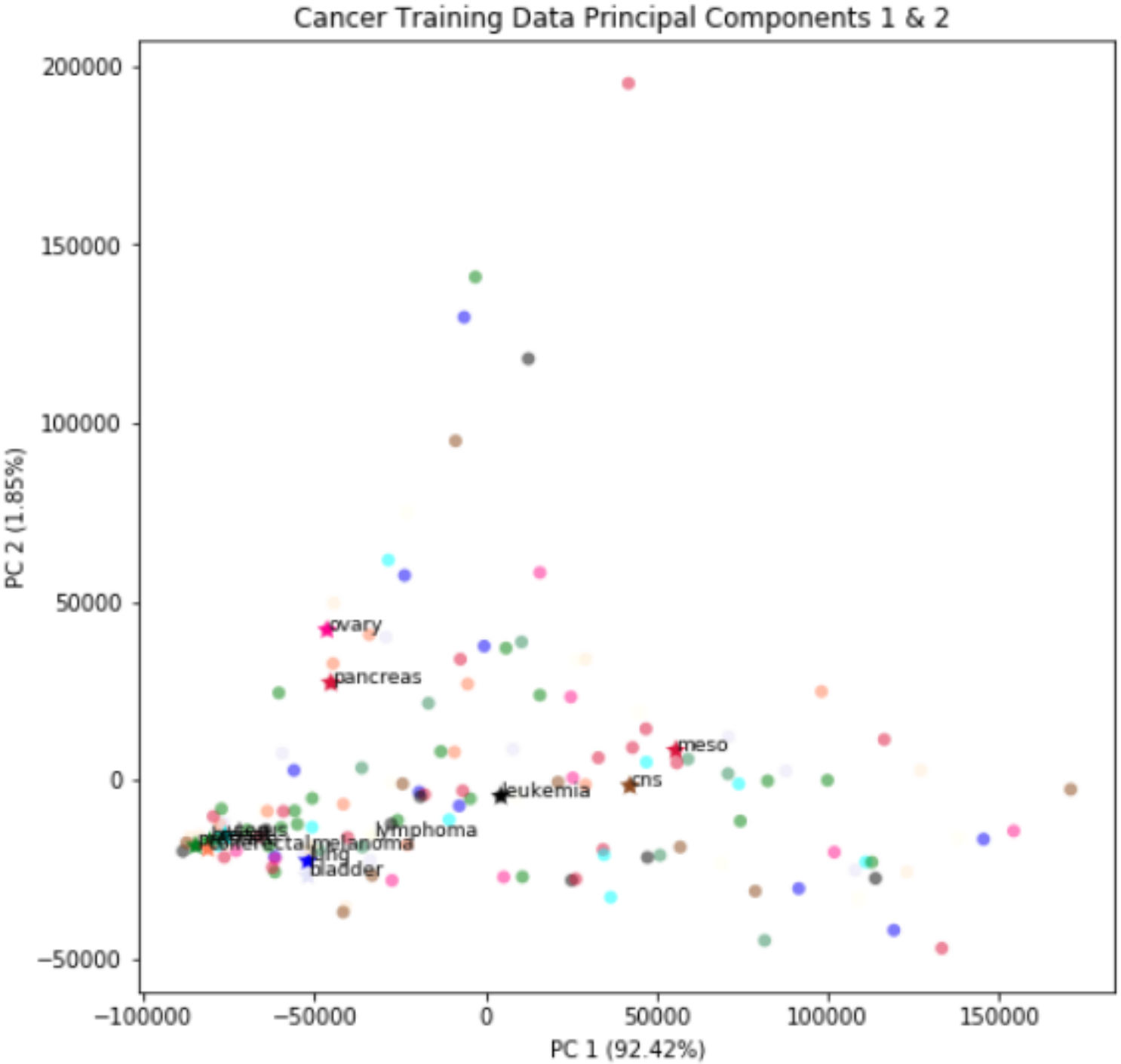

SOM Yeast Bicluster Data
Distancia Interneuronal
Output Neurons

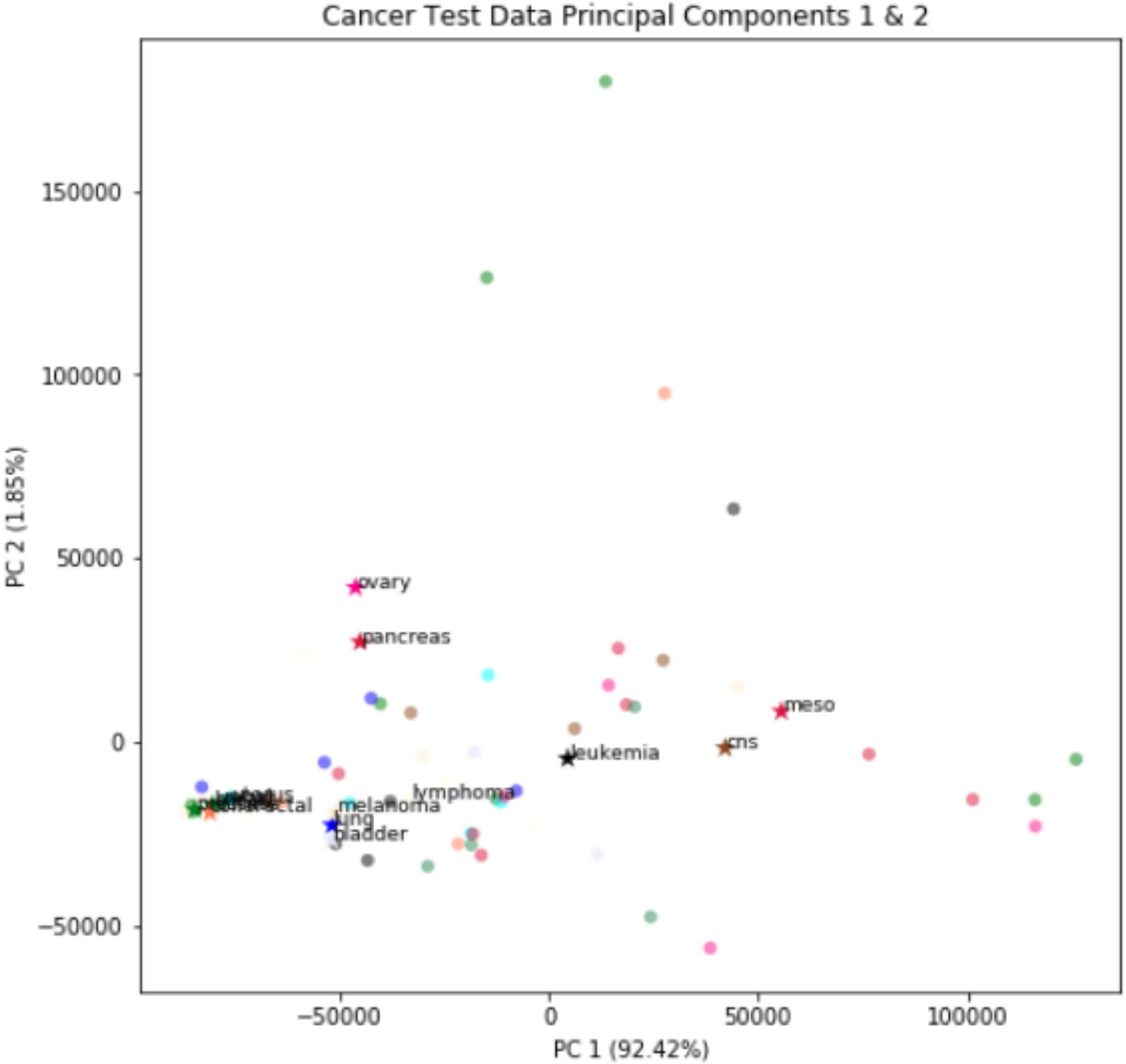

Human Data Prinicpal Components 1 & 2
PC2 5.25%
PC1 6.96%
1000
500
0
−500
−1000
−1500
−1000
−500
0
500
1000
1500


Human Bicluster Prinicpal Components 1 & 2
PC2 5.10%
PC1 7.14%
1000
500
0
−500
−1000
−1500
−1000
−500
0
500
1000
1500


SOM Human Data
Output Neurons
Distancia Interneuronal
0
2
4
6
8
10
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

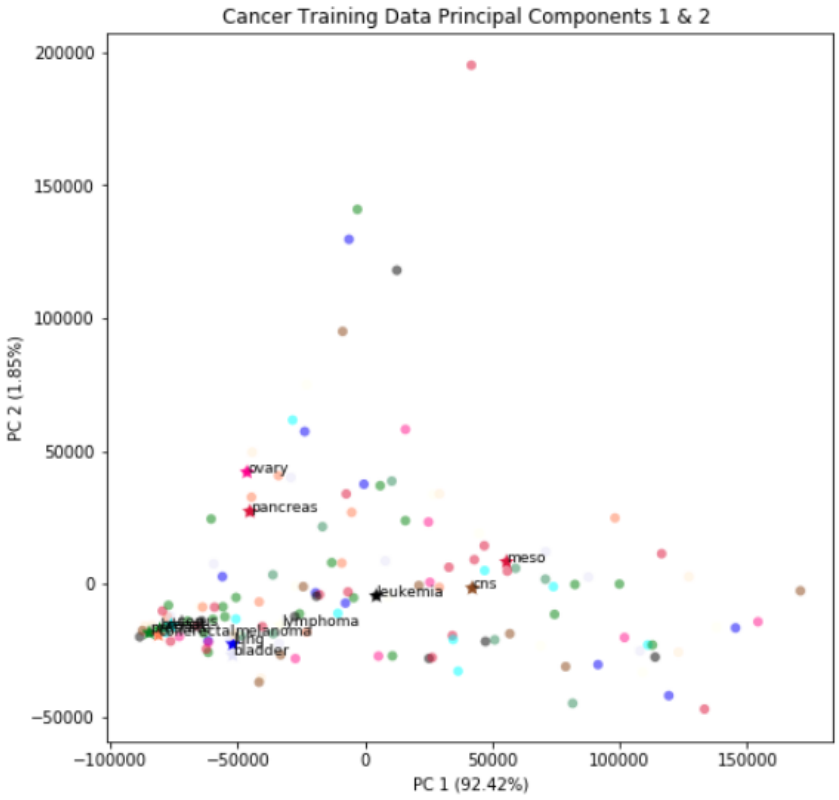

SOM Human Bicluster Data
Output Neurons
Distancia Interneuronal
0
2
4
6
8
10
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

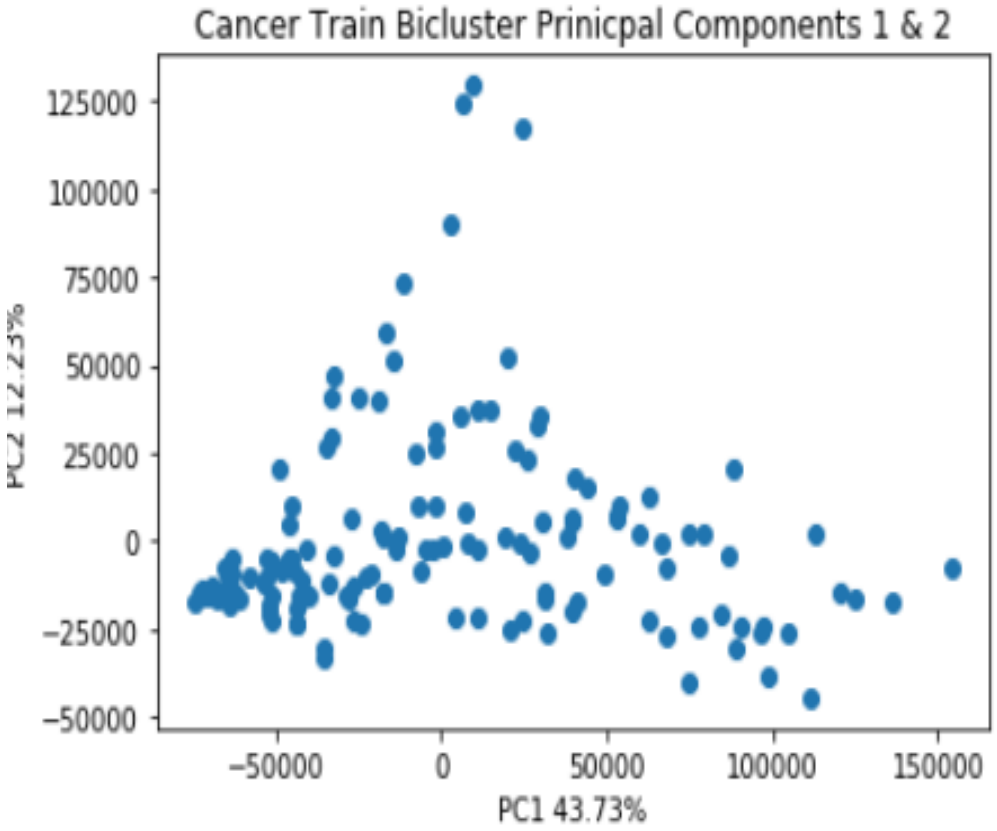

Cancer Training Data Principal Components 1 & 2
200000
150000
100000
50000
0
−50000
PC 2 (1.85%)
−100000
−50000
0
50000
100000
150000
PC 1 (92.42%)
ovary
pancreas
meso
cns
leukemia
lymphoma
melanoma
bladder


Cancer Test Data Principal Components 1 & 2
150000
100000
50000
0
−50000
PC 2 (1.85%)
ovary
pancreas
meso
cns
leukemia
lymphoma
melanoma
lung
bladder
−50000
0
50000
100000
PC 1 (92.42%)

Matriz de confusión datos de cáncer

labels	breast	prostate	lung	collerectal	lymphoma	bladder	melanoma	uterus	leukemia	renal	pancreas	ovary	meso	cns
breast	3	0	0	0	0	0	0	0	1	0	0	0	0	0
prostate	0	0	0	0	1	0	0	1	1	0	0	0	0	3
lung	0	0	0	0	1	0	0	1	1	0	0	0	0	1
collerectal	0	0	0	0	1	0	0	3	0	0	0	0	0	0
lymphoma	1	0	0	0	1	0	0	0	1	0	0	0	0	3
bladder	3	0	0	0	0	0	0	0	0	0	0	0	0	0
melanoma	1	0	0	0	1	0	0	0	0	0	0	0	0	0
uterus	1	0	0	0	0	0	0	1	0	0	0	0	0	0
leukemia	0	0	0	0	0	0	0	0	6	0	0	0	0	0
renal	1	0	0	0	0	0	0	0	2	0	0	0	0	0
pancreas	2	0	0	0	0	0	0	0	1	0	0	0	0	0
ovary	2	0	0	0	0	0	0	1	0	0	0	0	0	1
meso	0	0	0	0	1	0	0	1	0	0	0	0	0	1
cns	0	0	0	0	0	0	0	0	0	0	0	0	0	4


Cancer Training Data Principal Components 1 & 2
PC 2 (1.85%)
PC 1 (92.42%)
200000
150000
100000
50000
0
−50000
−100000
−50000
0
50000
100000
150000
ovary
pancreas
meso
leukemia
cns
lymphoma
melanoma
bladder


Cancer Train Bicluster Prinicpal Components 1 & 2
PC2 12.23%
PC1 43.73%
125000
100000
75000
50000
25000
0
−25000
−50000
−50000
0
50000
100000
150000

Conclusiones y Recomendaciones

- Realizamos la implementación y validación del algoritmo genético multi-objetivo: **MOEA/D** para la identificación de biclusters de interés en datos de microarreglos
- Utilizando las técnicas estadísticas: análisis de componentes principales (**PCA**) y mapa auto-organizado (**SOM**), se realizó un estudio comparativo de los datos originales y los biclusters encontrados mediante nuestra implementación del MOEA/D. Los biclusters encontrados contienen datos de interés.
- Recomendamos para trabajos futuros validar los biclusters obtenidos, para ver si forman conjuntos válidos de expresión, respecto a los reportados por la **comunidad biológica**, a través de múltiples conjuntos de

CICESE | Maestría en Ciencias de la Computación | Reconocimiento de Patrones

Biclustering para datos de Expresión Genética utilizando Optimización Multiobjetivo MOEA/D y posterior análisis mediante SOM y PCA.

Estudiantes:

- César Miguel Valdez Córdova
- Luis Enrique García Hernández