

Inteligencia Artificial aplicada a la clasificación de tumores mediante transcriptómica

Lic. Claudio Schuster

Transcriptómica

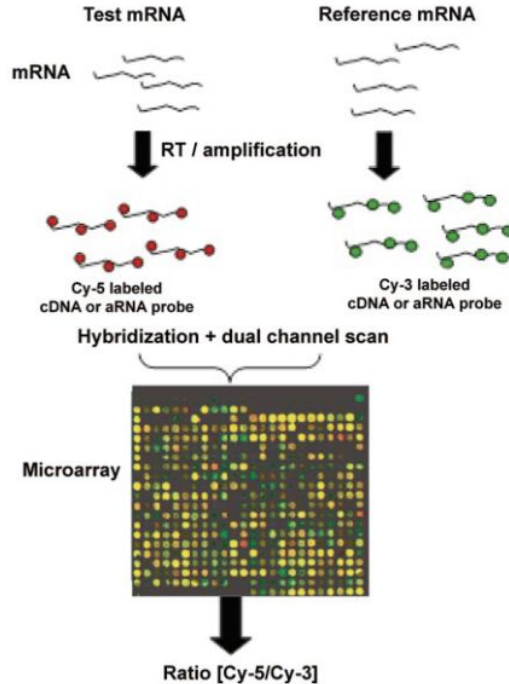
- Si la genómica se enfoca en la totalidad del ADN de un organismo (**genoma**), la transcriptómica tiene por objeto la totalidad del ARN (**transcriptoma**)
- El transcriptoma varía según el tipo celular, tiempo, patología, etc...
- Es un indicador de expresión génica. La cantidad de transcriptos (ARNm) de un gen puede usarse como estimador de su expresión
- Existen dos tecnologías para cuantificar expresión génica a escala “ómica”:
Microarrays y RNA-seq



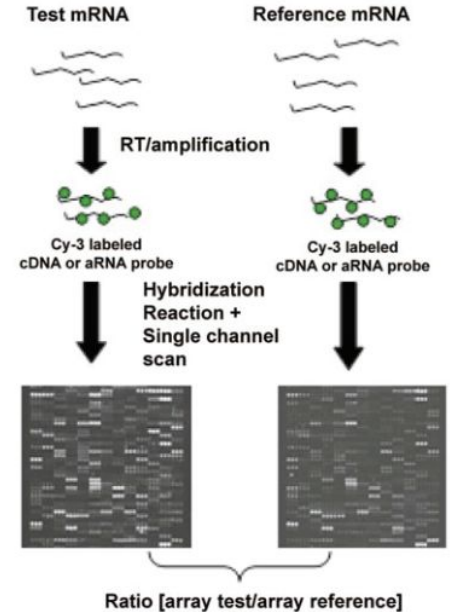
Microarrays

- Económicamente más accesible
- Permite muestrear más
- Produce datos más livianos y fáciles de procesar
- Tiene una cantidad fija de transcritos que puede detectar

A Two-color microarray hybridization



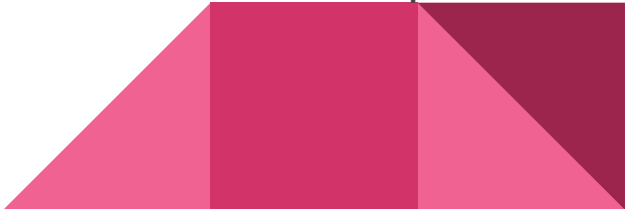
B One-color microarray hybridization



Los datos obtenidos

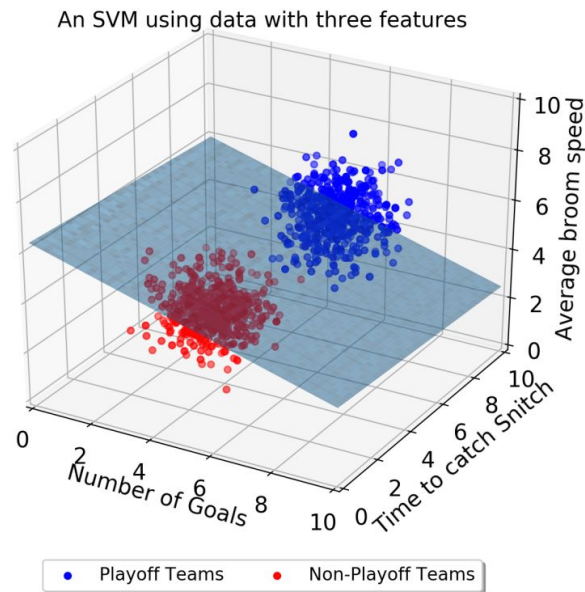
	A_A_WT_001	A_A_PTC_001	A_A_WT_002	A_A_PTC_002	A_A_WT_003	A_A_PTC_003
ENSG00000134569	3.77439452105285	10.5128798168438	5.18788630082566	12.1279778838265	6.50701255468133	9.40274181777523
ENSG00000160180	10.4198503507639	2.80577753035105	10.6129868659774	0.347067589006933	11.4492438104915	10.262335396549
ENSG00000123700	4.96284102293154	8.88387218074326	3.97176877013184	6.74073654517268	4.2031773841926	8.81789531864204
ENSG00000074706	8.50398468868402	1.97801053587308	8.78257973183492	1.89525473640381	9.43265358209412	7.8386008468766
ENSG00000197635	3.29445845997672	9.68449642756753	3.15549697203022	8.73148966370866	3.52472492094622	5.9692094733428
ENSG00000119655	11.809898467384	13.5258781574791	12.1012711979886	13.9168327314678	12.5104842659639	13.3007801774535
ENSG00000104723	4.02134144002743	7.15591046677727	4.03095609680849	7.33263968717127	3.96798269431761	4.88145502770426
ENSG00000166589	5.67509759364721	1.01993479700193	5.80570514421692	1.01993479700193	6.97998399096265	5.98891114160019
ENSG00000066382	8.30900444076468	2.94635662894733	8.18528682403489	3.06204831884359	7.82442692460574	5.98296995410844
ENSG00000153246	6.2314293734935	3.67462193851505	6.24228485390479	2.66266282920651	5.92754401056228	5.37883436969818
ENSG00000062038	4.20335603316466	8.32399477662394	4.20335603316466	9.39416792239863	3.53197912562836	7.57942221301951
ENSG00000130226	5.17927285360722	2.57953085883122	5.18410711007135	1.6250597882616	5.28408631235922	4.88614835643706
ENSG00000091129	4.21922831676992	5.89675286144582	4.49905491783791	6.12881385168051	4.27280575081612	5.28838516825111
ENSG00000143171	1.93575712219952	5.56325303204601	1.93575712219952	7.31661794042408	1.99349994336145	5.73110241838162
ENSG00000150672	4.56089547311938	2.42025571096677	4.765190546806	2.42691941707421	4.07486650746252	3.43905180946937

Machine Learning

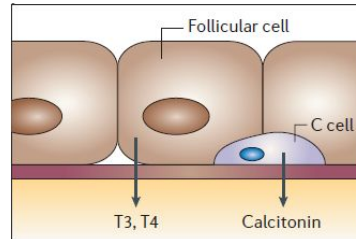
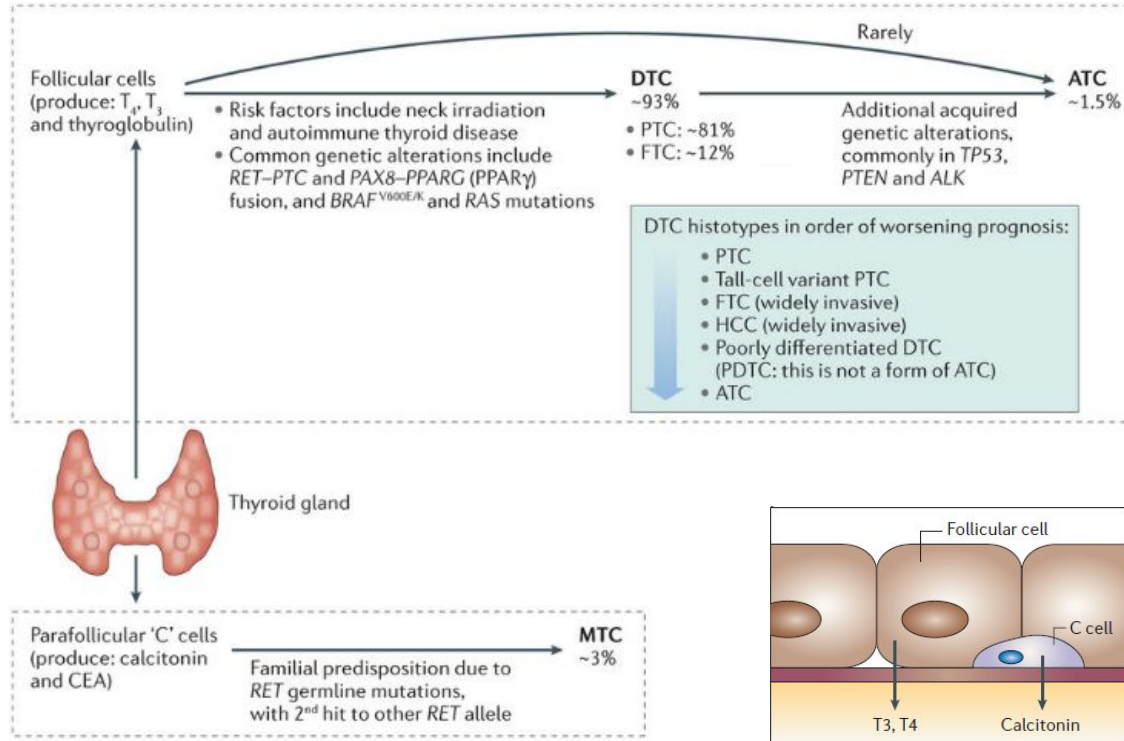
- Conjunto de técnicas computacionales que permiten aprender patrones y tomar decisiones sin intervención humana directa
 - Son una forma de Inteligencia Artificial
 - Según el tipo de datos que se disponga, pueden usarse para inferir el valor de una variable respuesta cuantitativa (regresión), clasificar o agrupar y detectar patrones en los datos
 - Los modelos de ML se entrenan con un conjunto de datos y se testean con otro distinto con el que se pone a prueba la capacidad del modelo para clasificar o inferir el valor de una VR
- 

Support Vector Machine (SVM)

- Es un tipo de modelo que permite clasificar de forma binaria
- Busca encontrar un hiperplano de N-1 dimensiones que separe de forma eficiente los datos

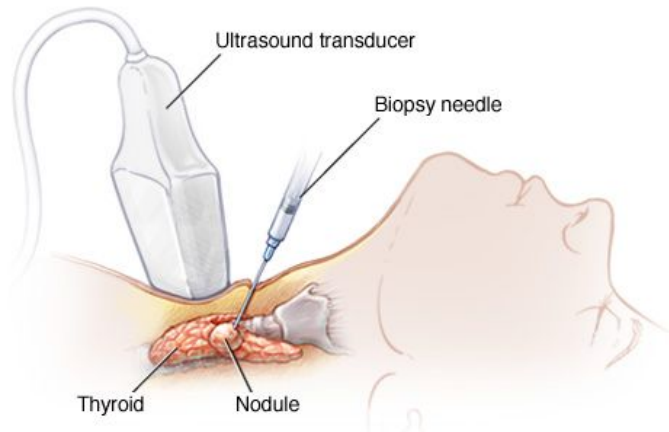


Cáncer de Tiroides



Cáncer de Tiroides

Bethesda category	Cytopathologic category	Approximate expected frequency	Malignancy rate	Suggested treatment (Prior to availability of molecular testing)
I	Non-diagnostic/Inadequate	5-11%	1-4%	Repeat FNA
II	Benign	55-74%	0-3%	US follow-up
III	Atypia/follicular lesion of undetermined significance	5-15%	5-15%	Repeat FNA or US follow-up or Lobectomy
IV	Follicular neoplasm/suspicious for FN	2-25%	15-30%	Lobectomy
V	Suspicious for malignancy	1-6%	60-75%	Lobectomy or Thyroidectomy
VI	Malignant	2-5%	97-99%	Near-total thyroidectomy



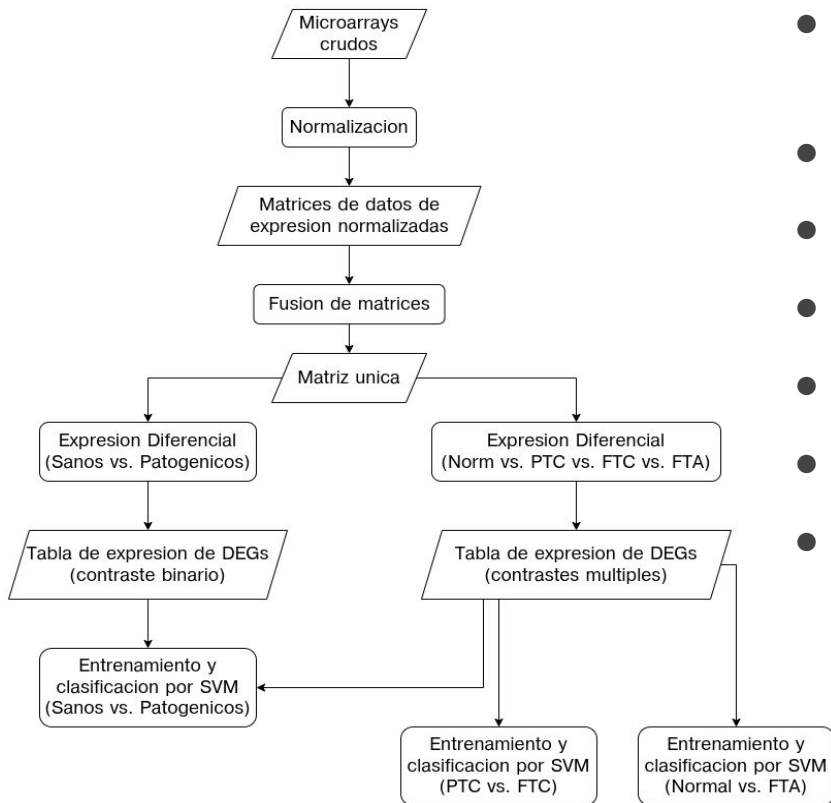
Objetivos

- Reunir múltiples experimentos de microarray para tejido tiroideo normal, PTC, FTC y FTA
- Obtener genes diferencialmente expresados para entrenar modelos de ML
- Entrenar un modelo de SVM que separe FTA y Normal de PTC y FTC
- Entrenar dos modelos de SVM para separar FTA de Normal y PTC de FTC

Transcriptómica? Para qué?



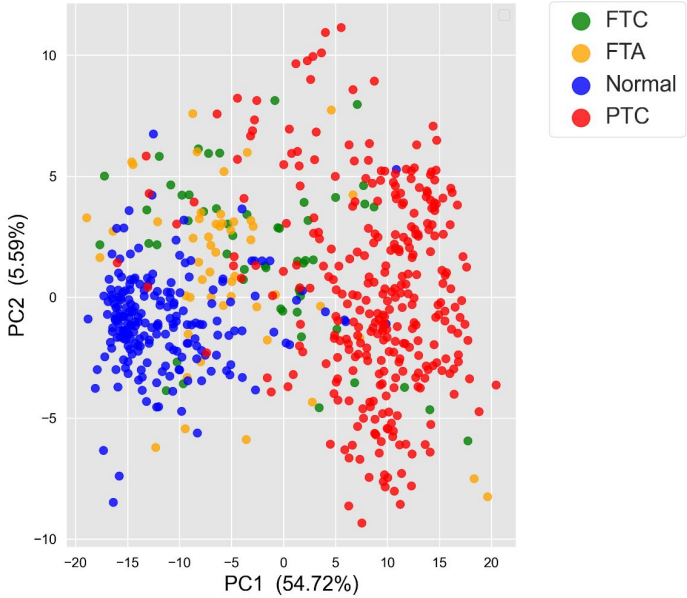
Métodos



- Microarrays de plataforma Affymetrix obtenidos de NCBI-GEO
- 203 Normales, 310 PTCs, 53 FTCs y 51 FTAs
- Normalización por *R-gcrma*
- Expresión diferencial en *R-limma*
- Entrenamiento de modelos en *scikit-learn* (python3)
- K fold CV: K = 5
- Kernel: Rbf (Radial Basis Function)

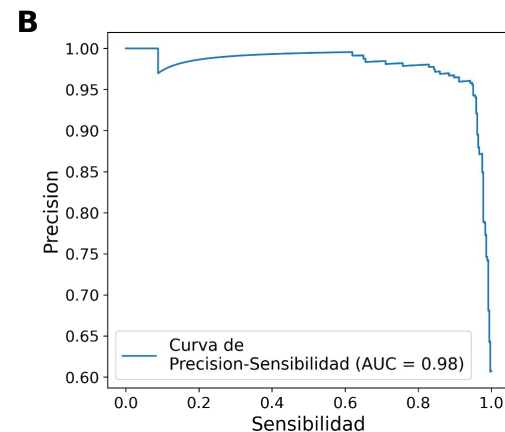
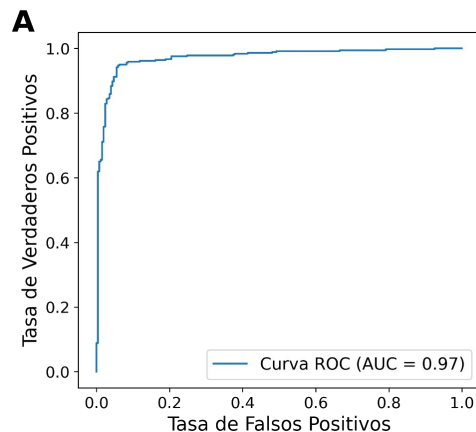
Resultados generales

		Modelo Binario	Modelo Sano	Modelo Carcinomas
Métrica	Error de Validación	6.15%	5.07%	7.59%
	Precisión	94.95%	90.85%	81.55%
	Sensibilidad	94.58%	83.12%	62.01%
	Especificidad	92.81%	97.9%	97.6%
	Valor Predictivo Negativo	92.29%	95.85%	93.76%
Hiperparámetro	C	2.4	4	3.3
	Gamma	1.69	0.04	0.44
	Número de Genes	40	445	110



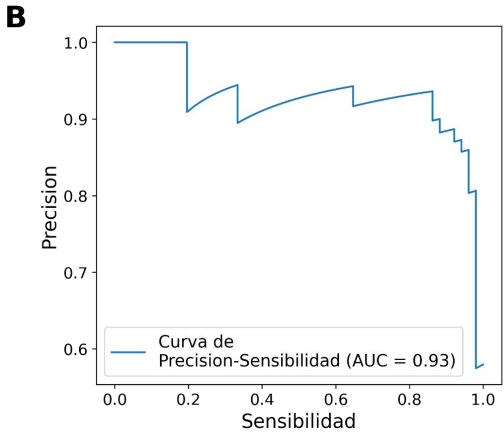
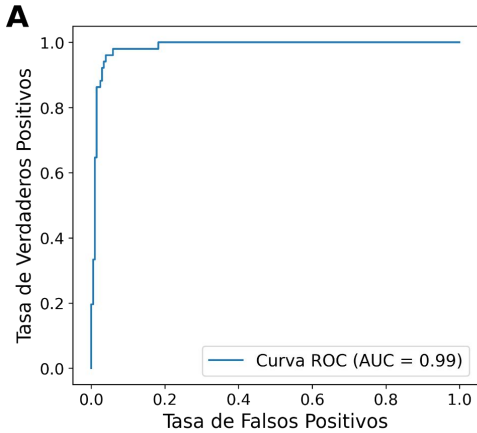
Modelo binario (Normal+FTA vs. PTC+FTC)

		Modelo Binario
Métrica	Error de Validación	6.15%
	Precisión	94.95%
	Sensibilidad	94.58%
	Especificidad	92.81%
	Valor Predictivo Negativo	92.29%
Hiperparámetro	C	2.4
	Gamma	1.69
	Número de Genes	40



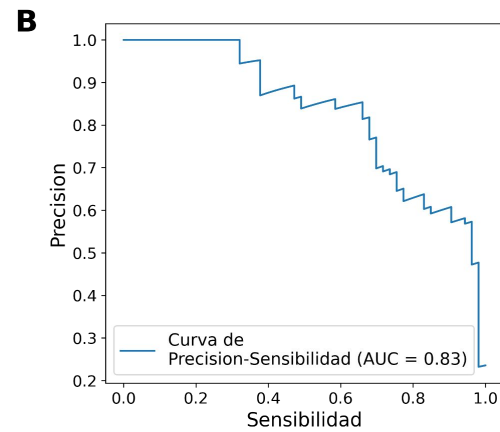
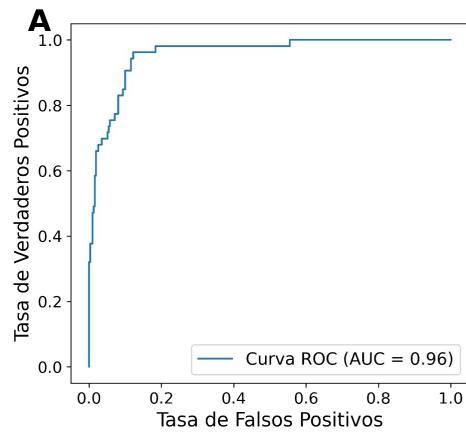
Normal vs. FTA

		Modelo Sano
Métrica	Error de Validación	5.07%
	Precisión	90.85%
	Sensibilidad	83.12%
	Especificidad	97.9%
	Valor Predictivo Negativo	95.85%
Hiperparámetro	C	4
	Gamma	0.04
	Número de Genes	445



PTC vs. FTC

		Modelo Carcinomas
Métrica	Error de Validación	7.59%
	Precisión	81.55%
	Sensibilidad	62.01%
	Especificidad	97.6%
	Valor Predictivo Negativo	93.76%
Hiperparámetro	C	3.3
	Gamma	0.44
	Número de Genes	110



Conclusiones

- Los modelos entrenados separan con mucha eficiencia los tipos de tumores analizados
- Presentan potencial de mejoría mediante un aumento del tamaño muestral y testeo con datos externos
- Se requiere aumentar especialmente el tamaño muestral de FTA y FTC para mejorar la capacidad de detección de estos



Perspectivas a futuro

- Aumentar el tamaño muestral de FTC y FTA mediante Bootstrap
- Obtener datos ajenos a los utilizados para entrenar los modelos y usarlos como datos de testeo
- Ampliar los modelos para que funcionen tanto con microarrays como RNA-seq



Agradecimientos



Dr. Marcelo Martí



Dr. Juan Pablo Nicola



Lic. Jonathan Zaiat

