Predicción de patogenicidad en SNPs usando Aprendizaje Automático

Tesis de Licenciatura en Ciencias de la Computación

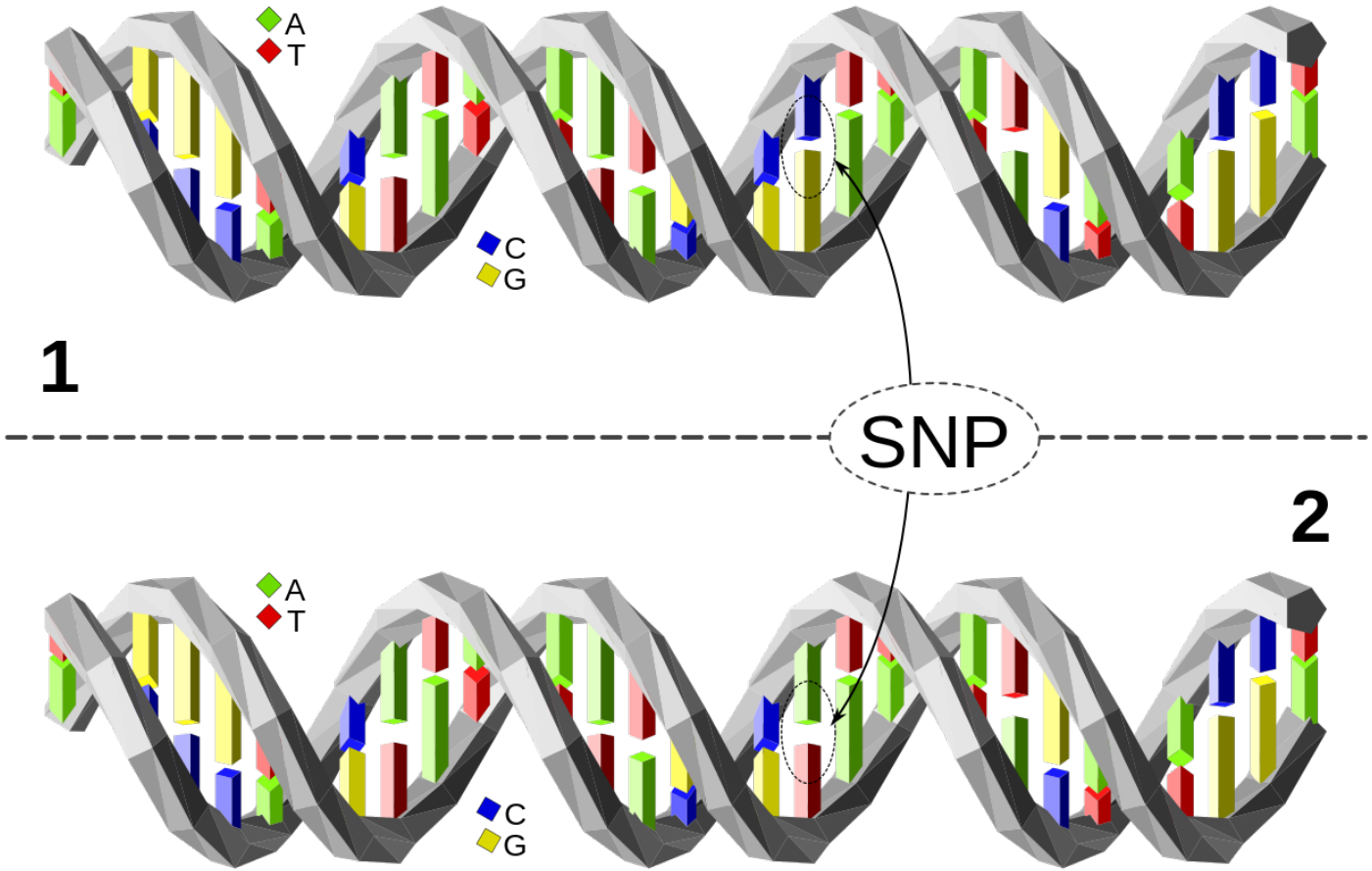
Martín Ezequiel Langberg
Directores: Ariel Berenstein y Pablo Turjanski
Jurado: Viviana Cotik y Marcelo Martí

Departamento de Computación, FCEyN, UBA

Motivación del trabajo

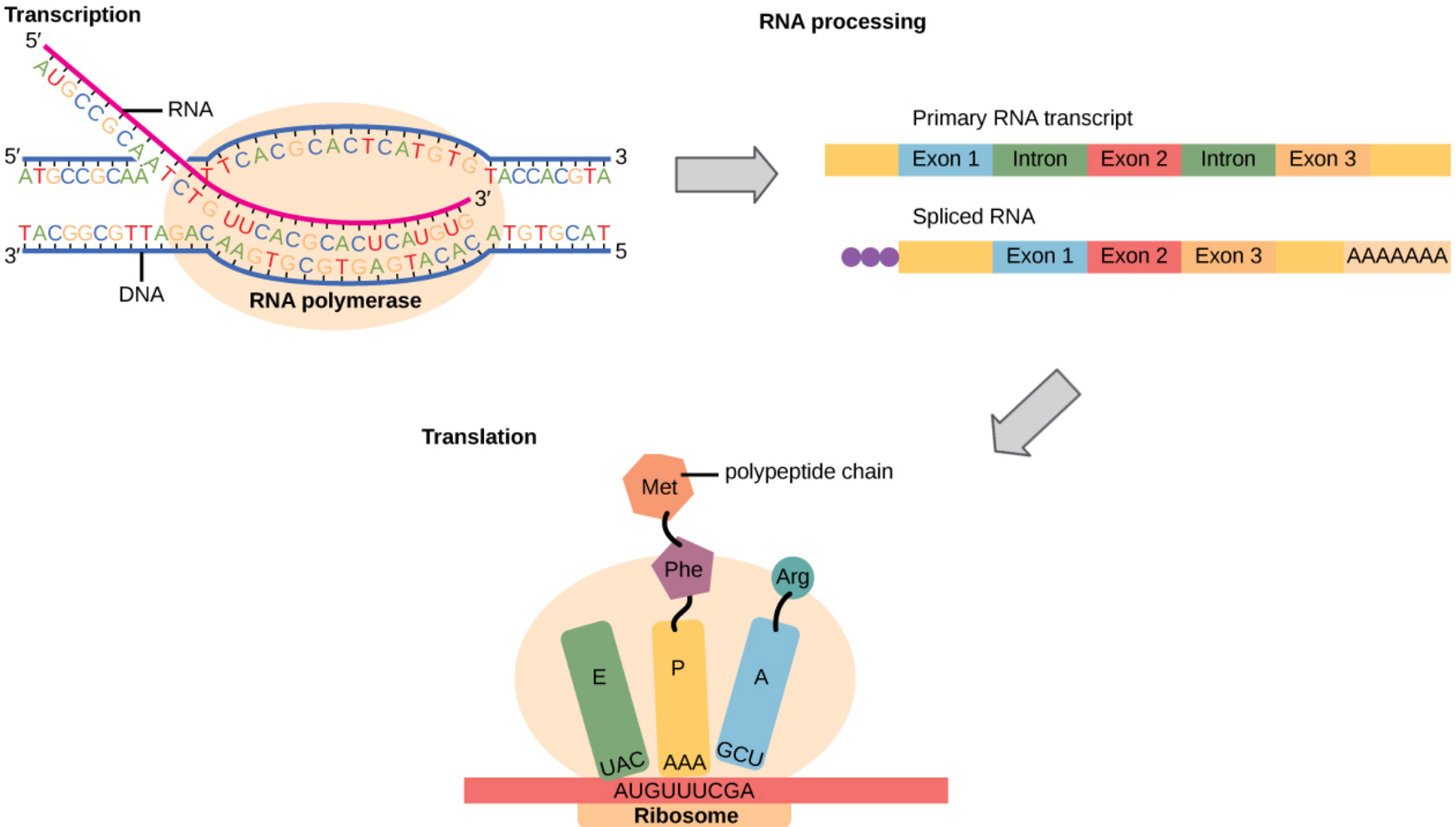
- Existen variantes en el ADN causantes de enfermedades (patogénicas)
- La detección de estas variantes es esencial para el avance de la medicina personalizada
- Queremos poder predecir la patogenicidad de un SNP en el ADN usando métodos computacionales

¿Qué son los SNPs?



Single Nucleotide Polymorphism (SNP)

Del ADN a las proteínas



Dogma central de la biología

¿Cómo se expresan los SNPs en el organismo?

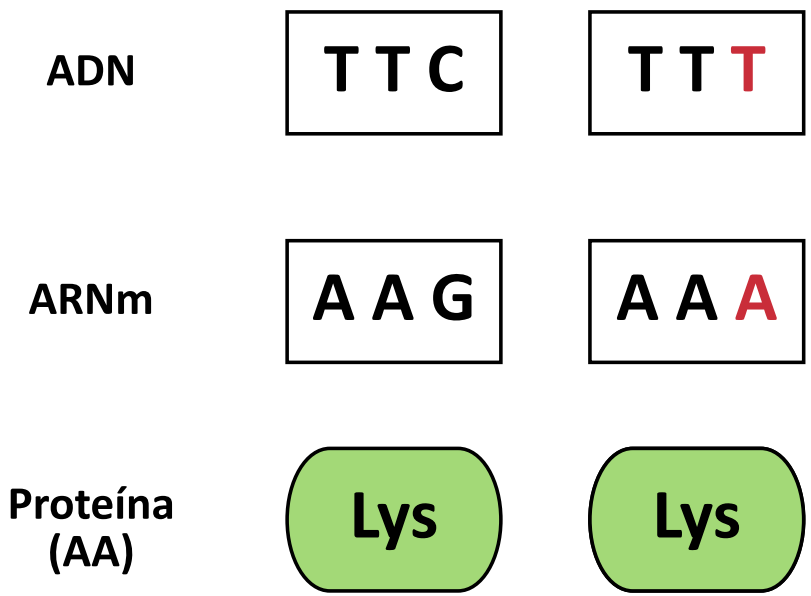
First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr **UAA Stop** **UAG Stop**	UGU, UGC: Cys **UGA Stop** UGG Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile **AUG Met**	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

Tabla de codones de ARN

Tipos de SNPs

Sustitución sinónima o *silent*

El cambio en el nucleótido no modifica el aminoácido

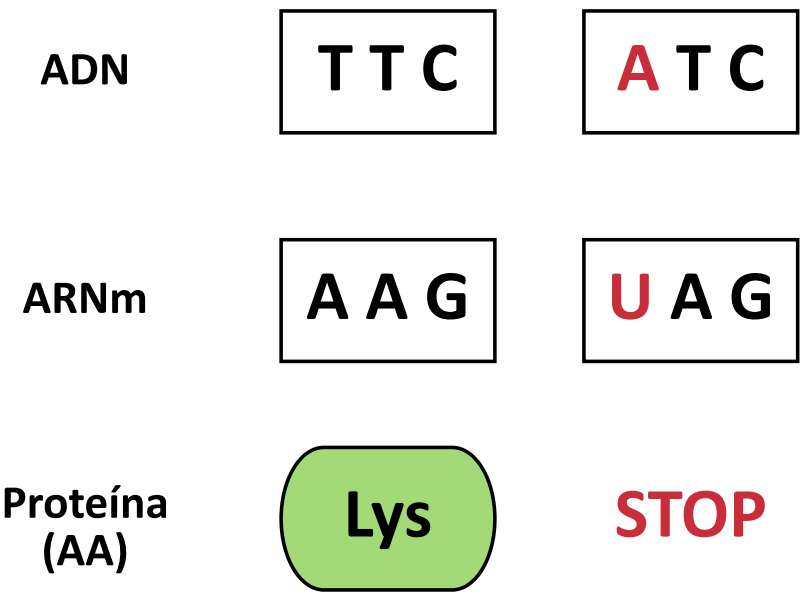


Sustitución *silent*

Tipos de SNPs

Sustituciones no sinónimas

Nonsense: Generan un codón de terminación o *stop*

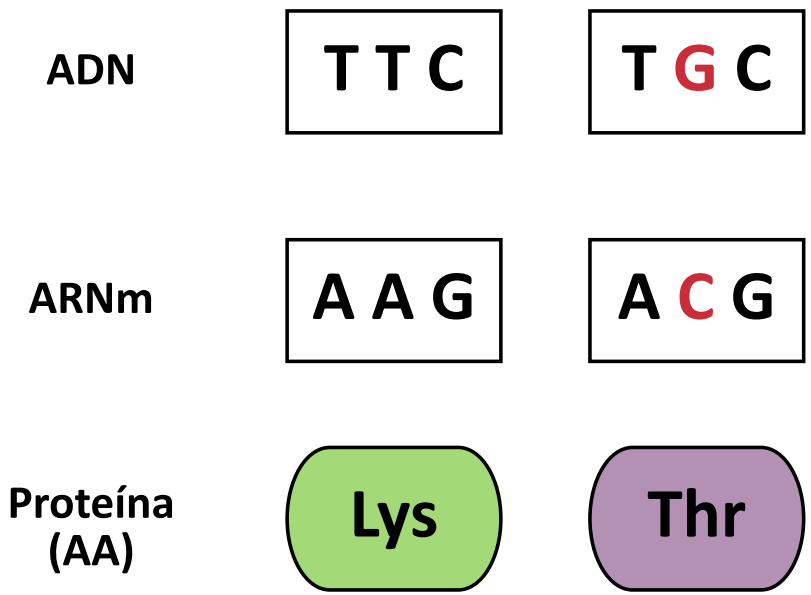


Sustitución *nonsense*

Tipos de SNPs

Sustituciones no sinónimas

Missense: Generan un cambio de aminoácido en la proteína



Sustitución *missense*

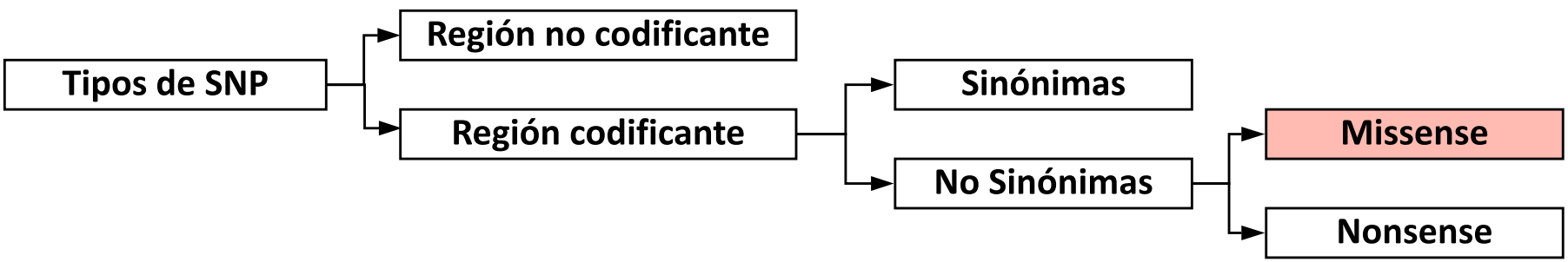
Foco de estudio: Variantes *missense*

Sustitución sinónima o *silent*

- El cambio en el nucleótido no modifica el aminoácido

Sustituciones no sinónimas

- Nonsense: Generan un codón de terminación o *stop*
- Missense: Generan un cambio de aminoácido en la proteína



Tipos de SNPs

Bases de datos biológicas

- Existen bases de datos biológicas que registran patogenicidad de variantes:
  - Clinvar (pública): Variantes de distinto nivel de confianza
  - **Humsavar (pública): Solamente variantes missense**
  - HGMD (privada)

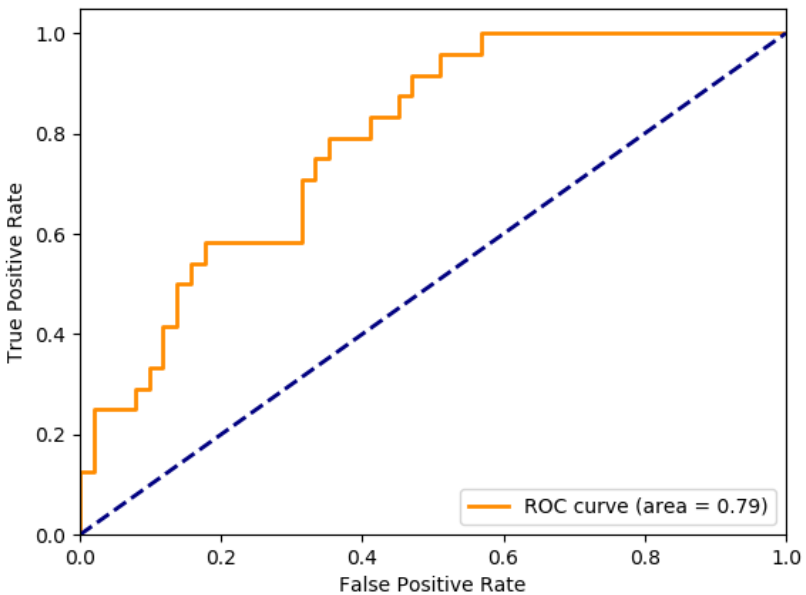
Swiss Prot AC	AA change	Type of variant	dbSNP
Q9NPC4	p.Pro251Leu	Polymorphism	rs28940571
Q9NPC4	p.Gln163Arg	Polymorphism	rs28915383
Q9NPC4	p.Ala218Asp	Polymorphism	rs2246945
Q9NRG9	p.His160Arg	Disease	-
Q9NRG9	p.Ser263Pro	Disease	rs121918550

Selección de columnas de tabla Humsavar (extracto)

Enfoque computacional: un problema de clasificación

- Trabajos previos:
  - VEST (Carter et al., 2013)
  - FATHMM-MKL (Shihab et al., 2015)
  - REVEL (Ioannidis et al., 2016)
  - VarQ (Santiago Moreno)
- Aprendizaje automático supervisado
- Dimensiones estructurales, físico-químicas de las proteínas, genómica
- Análisis de importancia de los features

Principal métrica de desempeño: AUC (Area bajo la curva)



Curva ROC

Rango de valores

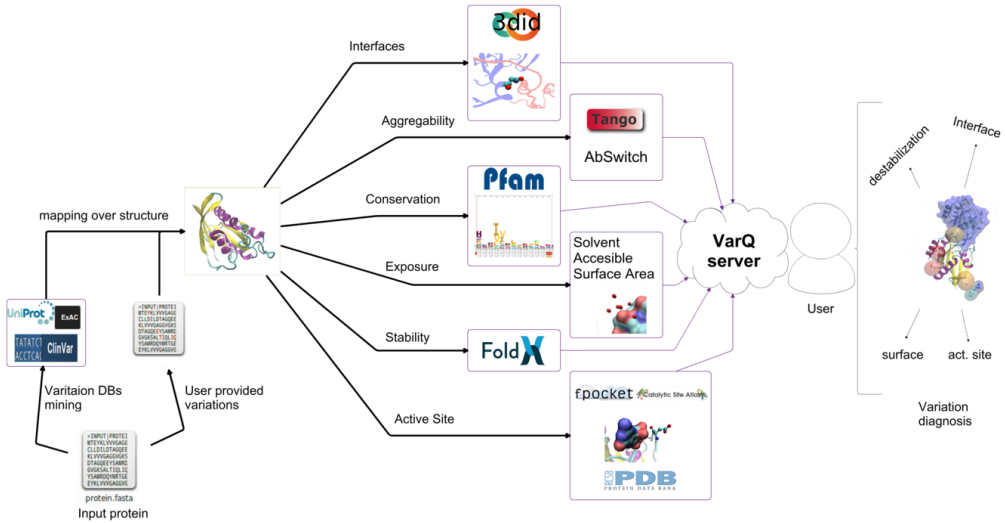
- AUC = 1: Predictor ideal
- AUC = 0.5: Predictor *random*

Principales ventajas

- Independiente del umbral de clasificación
- No es sensible a desbalances en los datos

¿Qué tan difícil es este problema?

Primer modelo: Propiedades estructurales usando VarQ



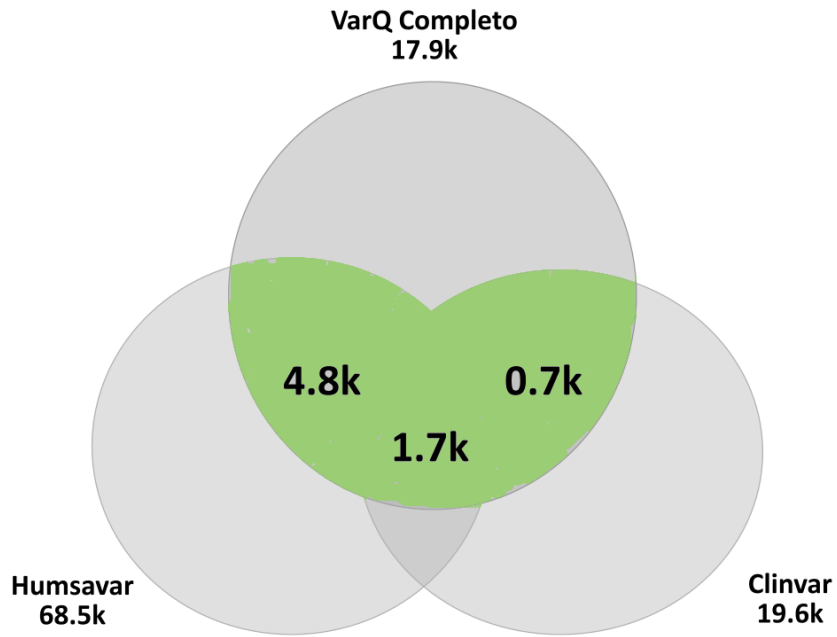
Pipeline de extracción de datos de VarQ

Features extraídos (cobertura)

- Variación de la energía (100 %)
- SASA (95 %)
- Porcentaje de SASA (95 %)
- B-Factor (95 %)
- Switchability (90 %)
- Aggregability (72 %)
- Conservación (37 %)
- Interfaz 3DID (100 %)
- Interfaz PDB (100 %)
- **Active Site (5 %)**

Filtrado de variantes del dataset VarQ

- Removimos variantes sin un status confirmado (*risk factor*, *likely benign*, *uncertain significance*)
- Aproximadamente 7.5k variantes: 72 % patogénicas, 28 % benignas
- Dataset VarQ Curado



Intersección del dataset VarQ usando Humsavar y Clinvar

Generación de modelos de aprendizaje automático

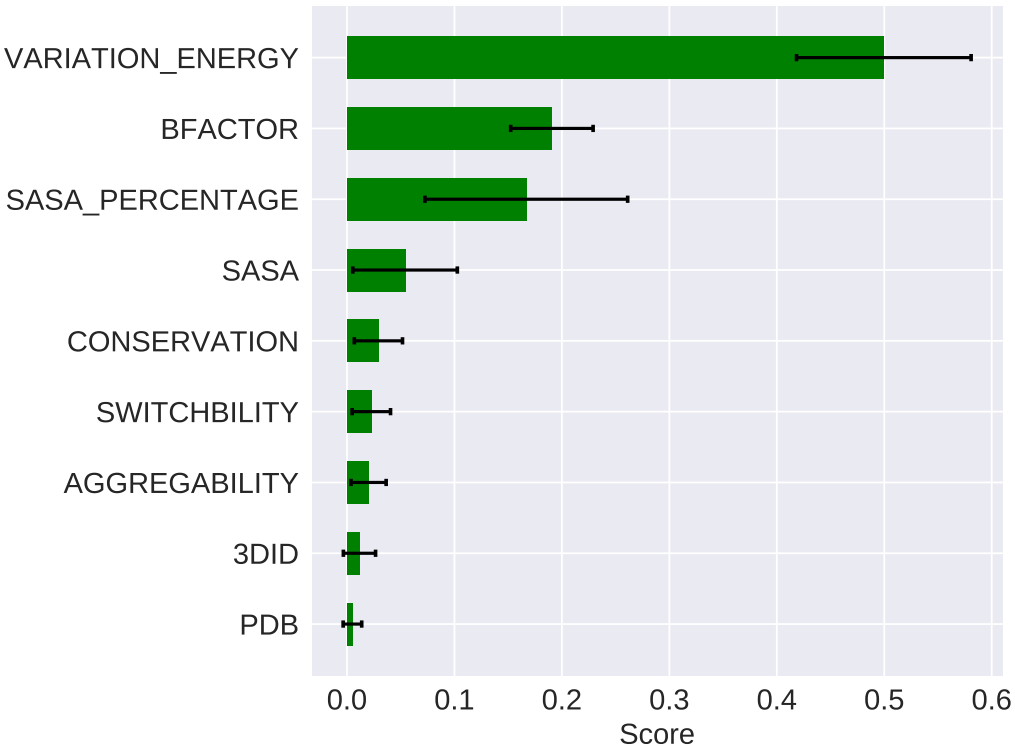
- Modelos clásicos usando `scikit-learn`
  - Support Vector Classifier (kernel radial)
  - Random Forest
  - Regresión logística
- Imputación de features nulos
  - Mediana para features continuos
  - Media para features categóricos (PDB, 3DID)
- Búsqueda de hiperparámetros usando *Grid-search* y validación cruzada
- 70 % dataset de entrenamiento, 30 % dataset de test

El modelo Random Forest obtuvo el mejor AUC

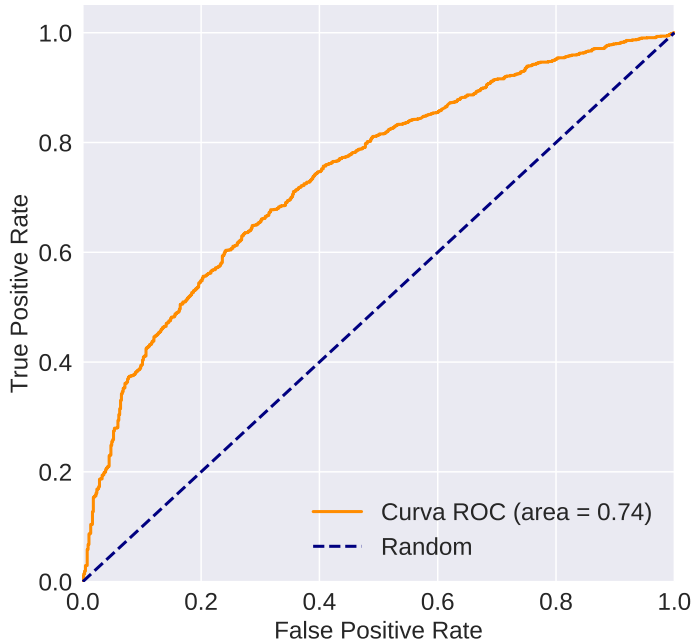
	SVC	LR	RF
Precisión	0.72	0.75	**0.77**
Recall	**1.00**	0.94	0.93
AUC	0.70	0.71	**0.74**
T_{fit}	2m 39s	**1.17s**	9.82s
T_{pred}	0.77s	**0.01s**	0.11s

- Hiperparamétros óptimos encontrados (RF):
  - Profundidad de árbol: 7
  - Estimadores: 100
  - Cantidad de variables por árbol: 4

La variación de la energía es el feature más importante



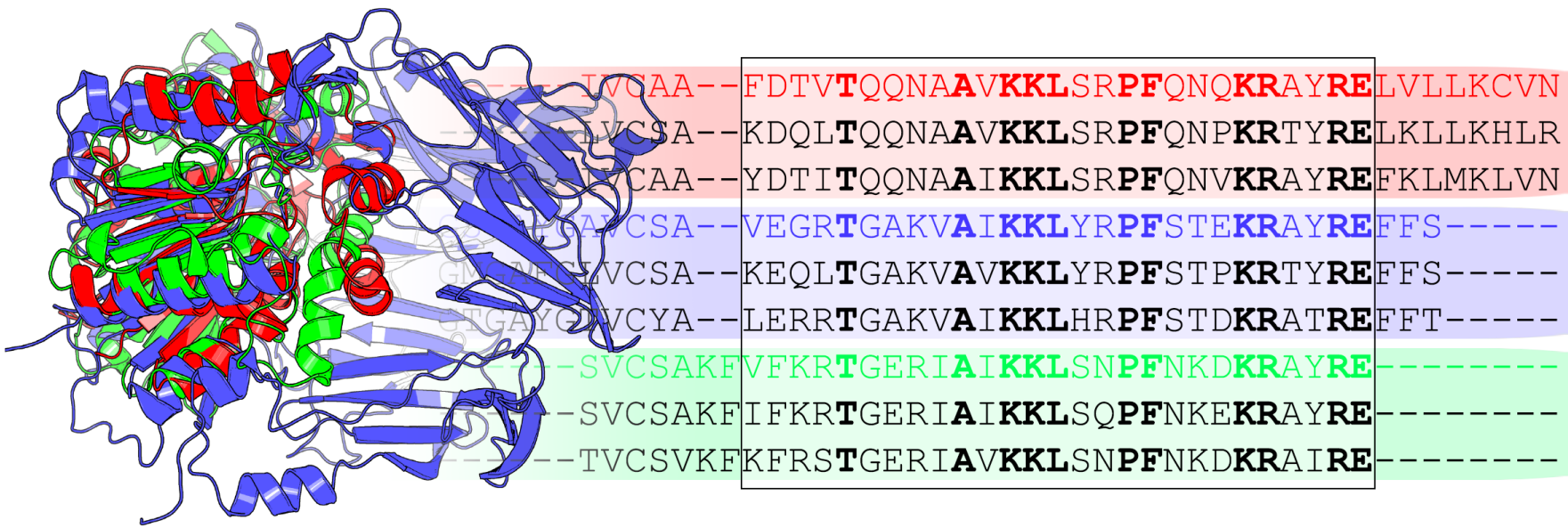
Importancia de features usando método estándar de `scikit-learn`



Curva ROC

- **AUC VarQ Curado: 0.74**
- AUC VEST: 0.84
- AUC FATHMM-MKL: 0.82
- AUC REVEL: 0.90

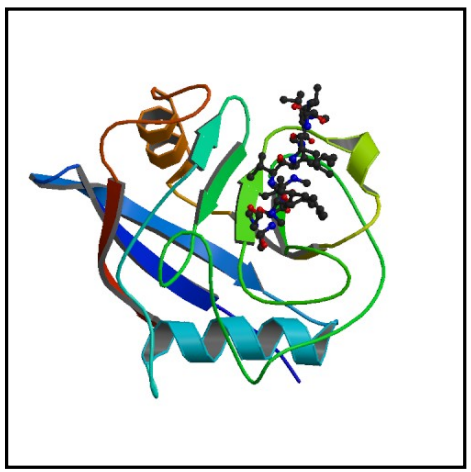
¿Cuál es el valor predictivo de las propiedades fisico-químicas de la proteína?

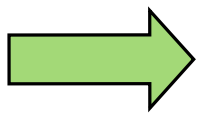


https://biokinet.belozersky.msu.ru/mustguseal

Modelo: Propiedades Físico-Químicas de la proteína

- **Usando únicamente la tabla Humsavar**:
  - Más de 68 mil variantes (aprox. x10 Varq!)
  - Status aportado por Humsavar
- Uniprot: Proteoma humano completo
- Nuevas fuentes de features:
  - ProtParam (Biopython)
  - SNVBox





```
>3ODL:A|PDBID|CHAIN|SEQUENCE
  MVNPTVFFDIAVDGEPLGRVSFEL
  FADKVPKTAENFRALSTGEKGFGY
  KGSCFHRIIPGFMCQGGDFTRHNG
  TGGKSIYGEKFEDENFILKHTGPG
  ILSMANAGPNTNGSQFFICTAKTE
  WLDGKHVVFGKVKEGMNIVEAMER
    FGSRNGKTSKKITIADCGQLE
```

Extracción de secuencia proteica (Ciclofilina A: P62937) en formato FASTA usando Uniprot

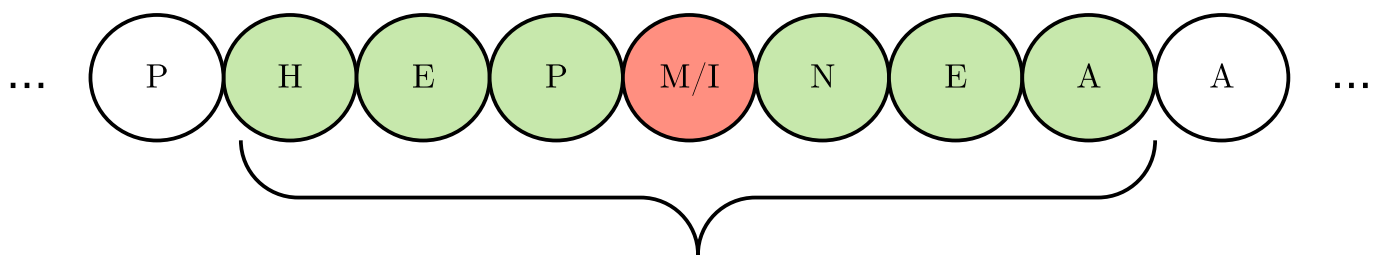
Generación de nuevos features usando ProtParam

Parámetros calculados

- Punto isoeléctrico
- Aromaticidad
- Índice de inestabilidad
- Flexibilidad
- Promedio de hidrofobicidad

Cambio en la variante

- Diferencia: $x_{var} - x$
- Log-ratio: $log(x_{var})/log(x)$



Ventana de cálculo del parámetro

Variables físico-químicas extraídas de SNVBox

Variables a nivel de aminoácido (considerando sustitución)

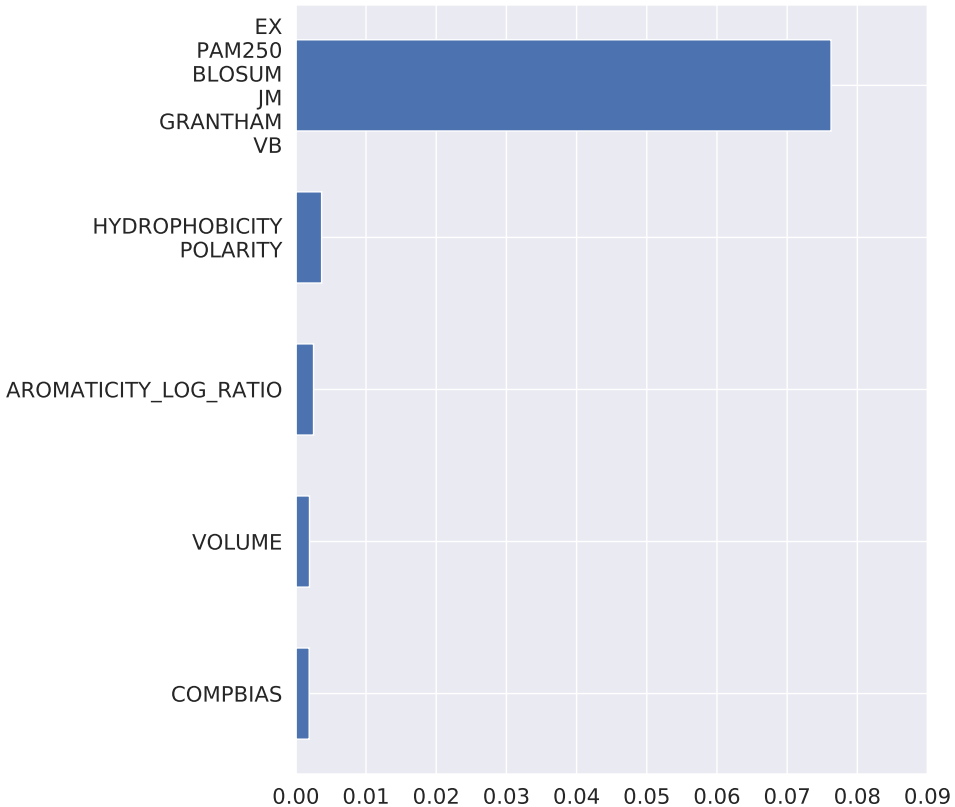
- Score BLOSUM, EX, GRANTHAM, PAM250, VB, JM
- Carga
- Volumen
- Polaridad
- Hidrofobia
- Transición

Variables a nivel de proteína (sin considerar sustitución)

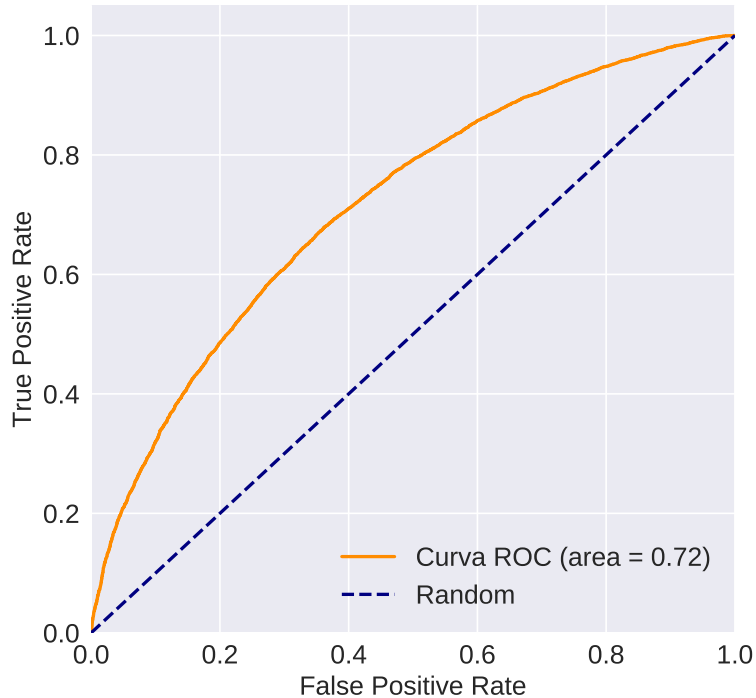
- BINDING: Sitio de unión
- ACTIVE_SITE: Sitio activo
- LIPID: Unión con un lípido
- METAL: Unión con un metal

- Base de datos generada en la Universidad Johns Hopkins

Las matrices de sustitución fueron las más relevantes



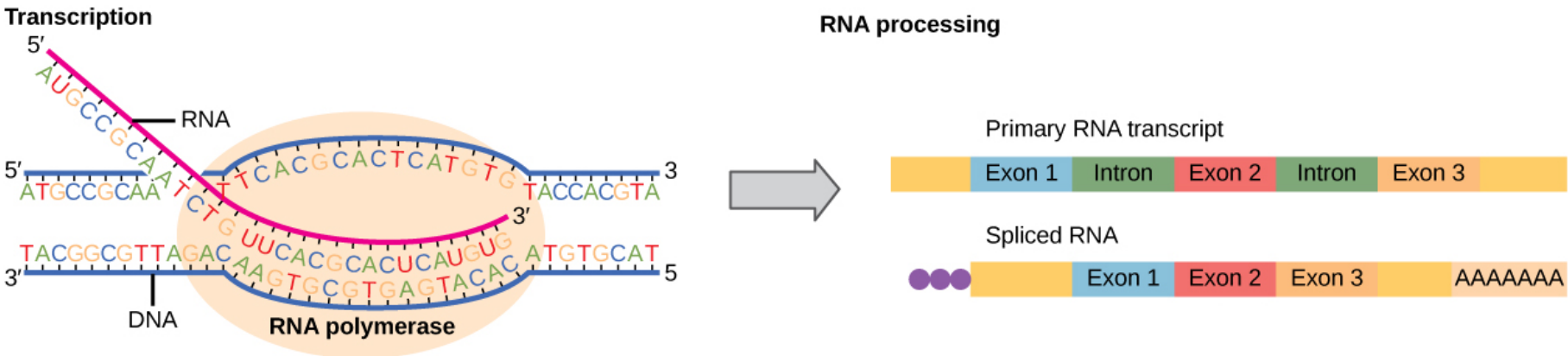
Importancia de features clusterizada usando `rfpimp`



Curva ROC

- **AUC Físico-Químico: 0.72**
- AUC VarQ: 0.74

¿Cuál es el valor predictivo de los features genómicos?



Modelo: Variables genómicas

- Identificador rsID: aproximadamente 55k variantes en Humsavar
  - 68 % variantes benignas
  - 32 % variantes patogénicas
- Fuentes de features:
  - SNVBox
  - dbSNP
  - Genome Browser (UCSC)

Explorador de variantes de dbSNP (`https://www.ncbi.nlm.nih.gov/snp`)

Features del modelo Genómico

Features de conservación genómica

- PhastCons a 46 vías (vertebrados)
- PhyloP a 46 vías (vertebrados)

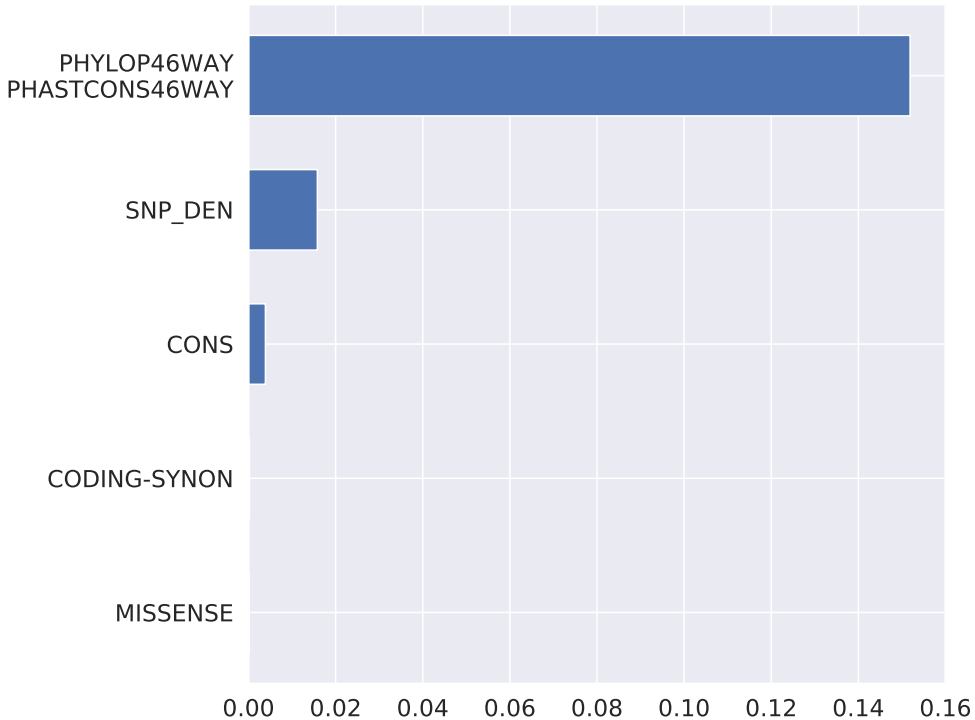
Features extraídas de SNVBox

- Conservación a nivel de exón
- Densidad de SNPs en HapMap
- Densidad de SNPs a nivel de exón

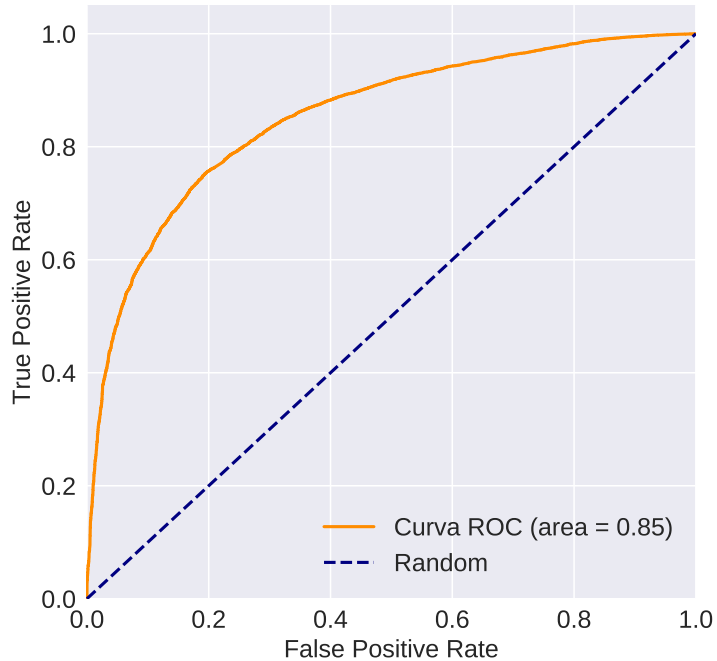
Features relativas a la clase funcional

- Missense
- Nonsense
- Intrón

La conservación genómica aportó un salto en el AUC



Importancia de features clusterizada usando `rfpimp`



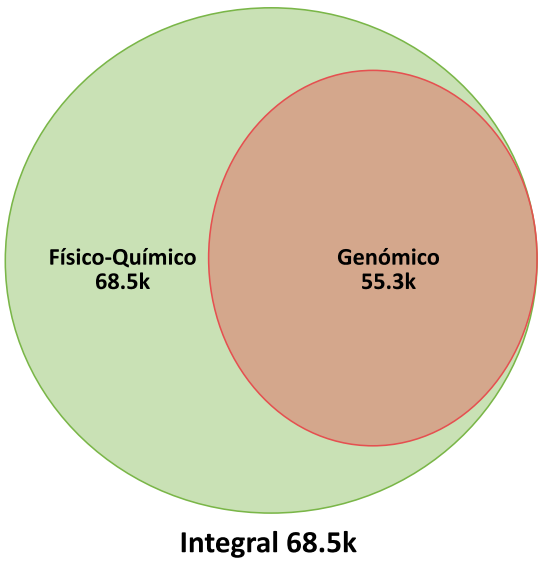
Curva ROC

- **AUC Genómico: 0.85**
- AUC Físico-Químico: 0.72
- AUC VarQ: 0.74

¿Podemos mejorar el modelo genómico integrando los features físico-químicos?

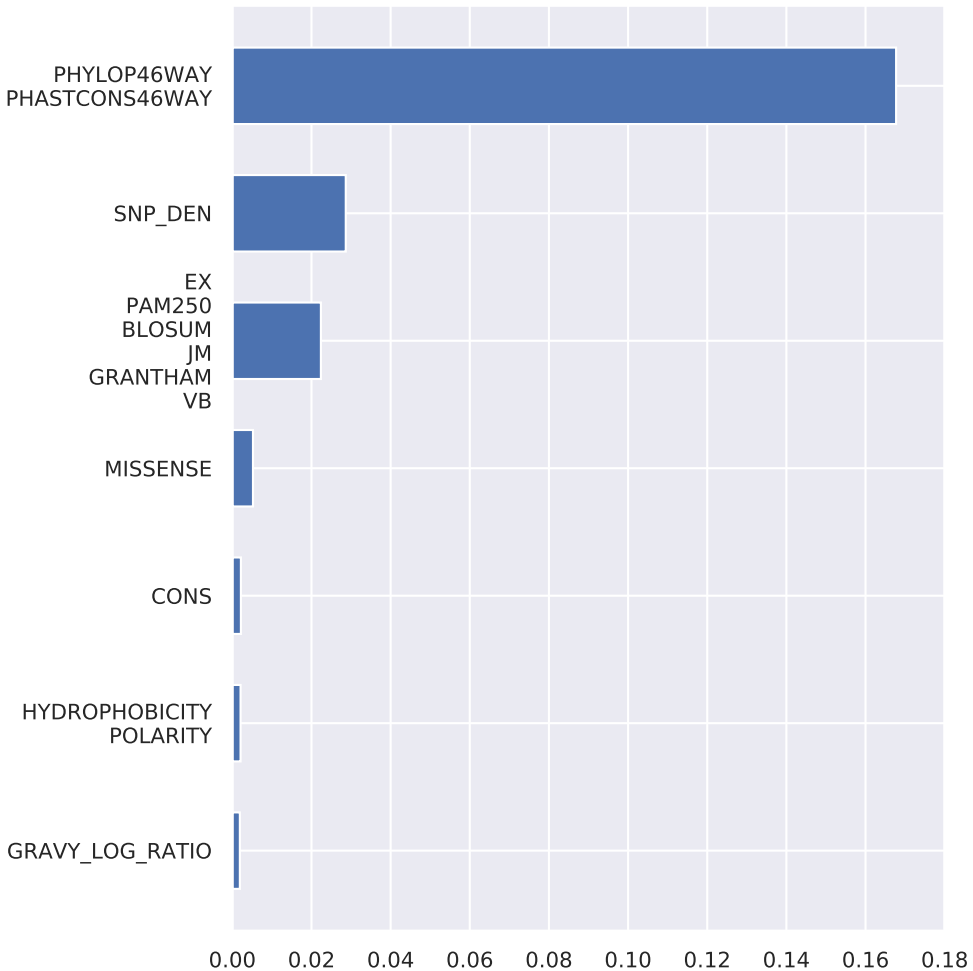
Integramos los features físico-químicos y genómicos

- Dataset Humsavar: 68 mil variantes
- Cobertura features genómicos: aprox. 80 %
- Cobertura features físico-químicos: misma que el dataset físico-químico
- Evaluamos un nuevo método de aprendizaje automático: XGBoost

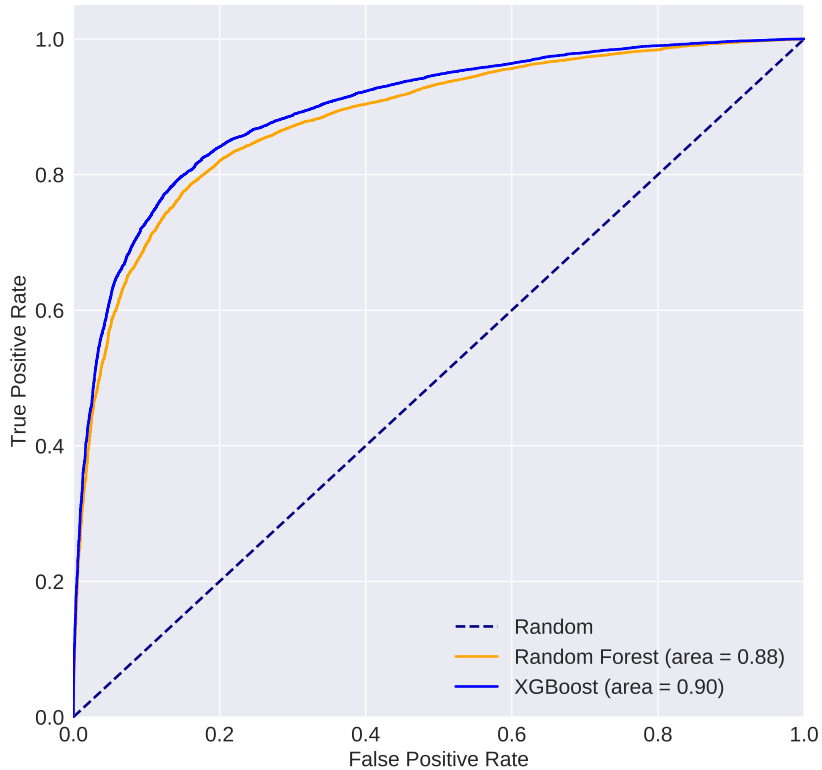


Unión de los datasets Físico-Químico y Genómico

XGBoost permitió alcanzar al mejor trabajo del área

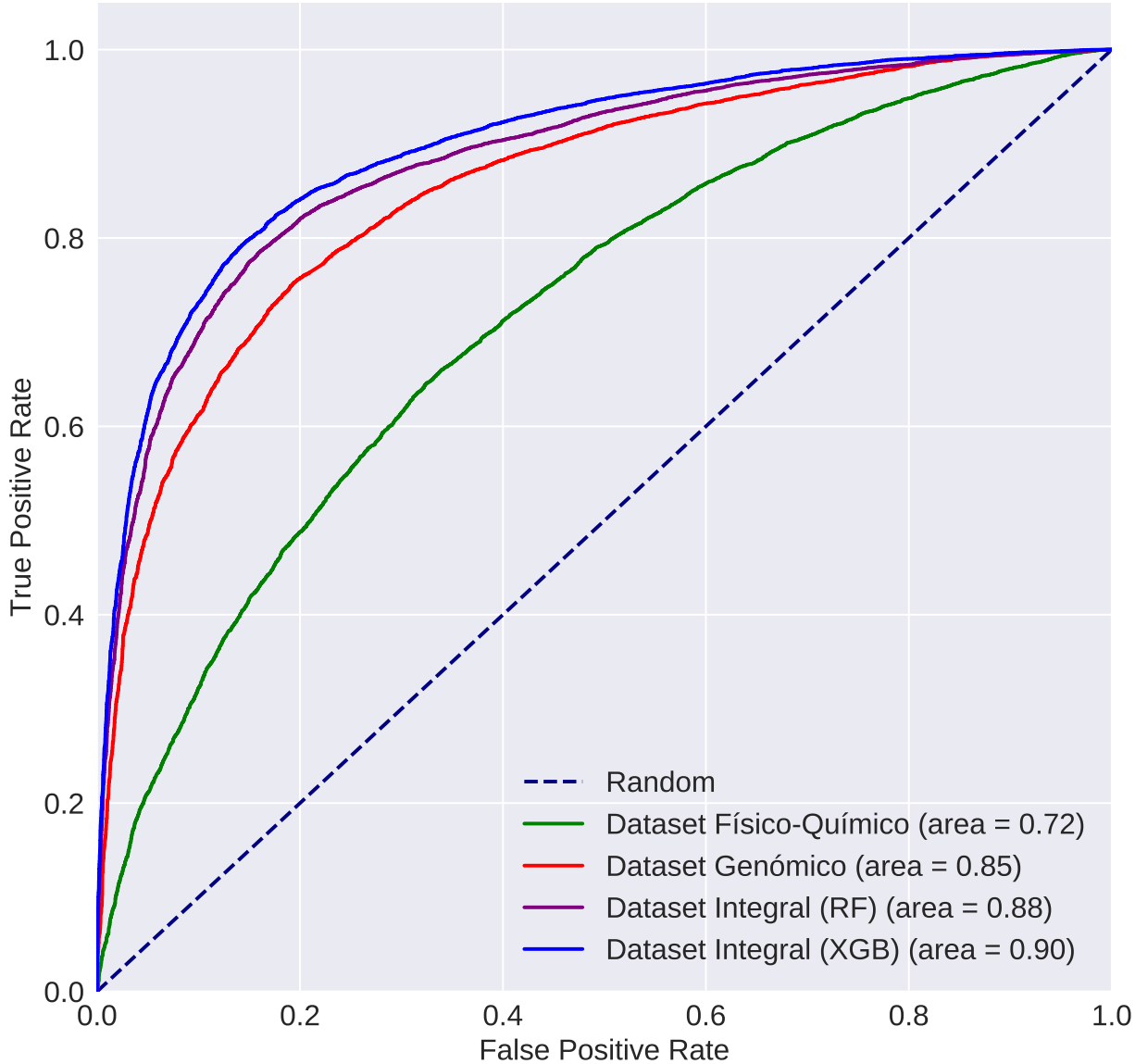


Importancia de features clusterizada usando rfpimp



Curva ROC

- **AUC Integral: 0.90**
- AUC VEST: 0.84
- AUC FATHMM-MKL: 0.82
- **AUC REVEL: 0.90**

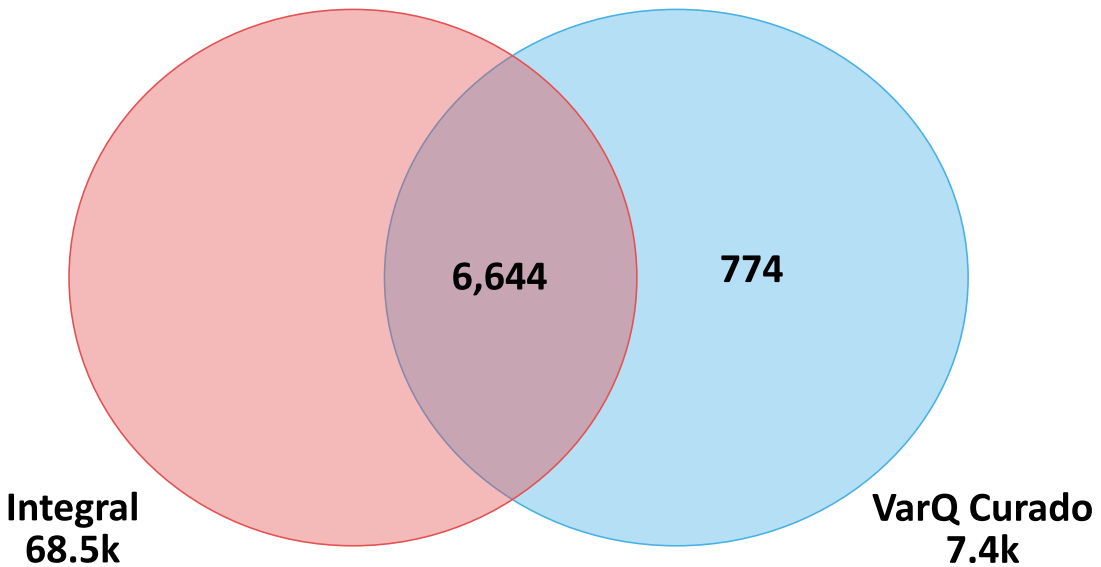


Desempeño de los modelos usando las variantes de la tabla Humsavar

¿Qué sucede al sumar estos features al dataset VarQ?

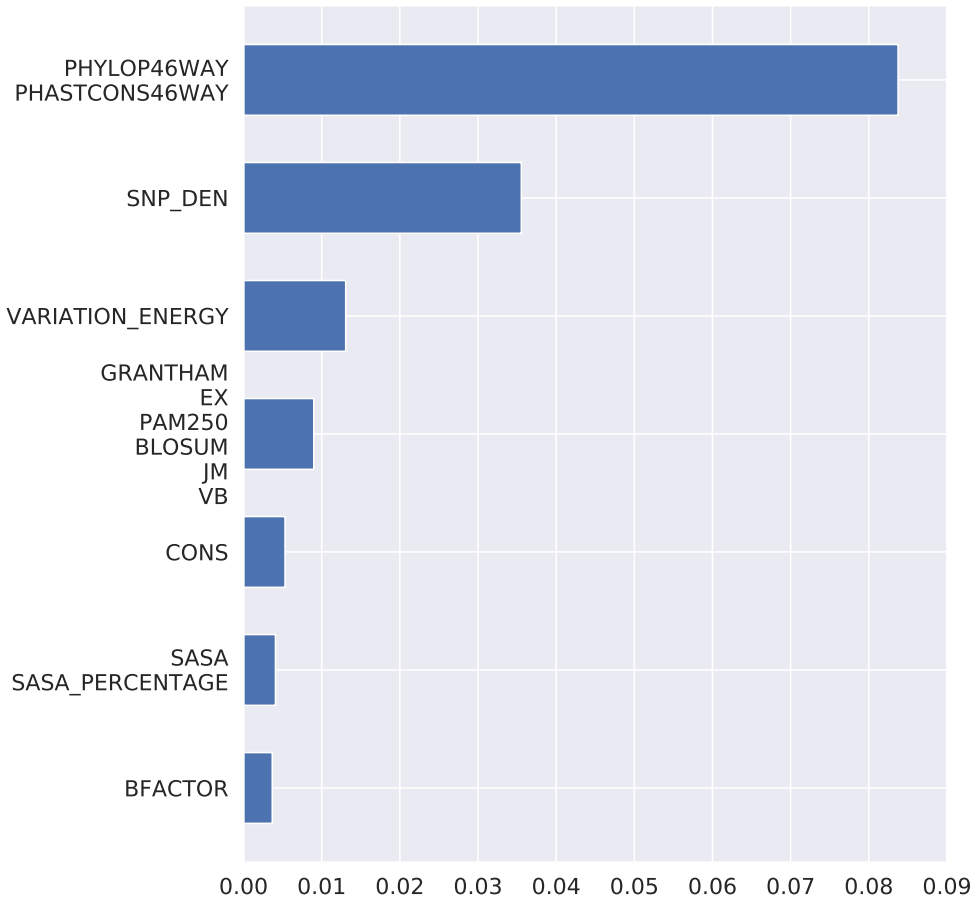
Sumamos las nuevas variables al dataset VarQ

- Unimos los features del dataset Integral al dataset VarQ Curado
- 72 features: 9 de VarQ + 63 de Integral
- 7.4k variantes: 72 % patogénicas, 28 % benignas

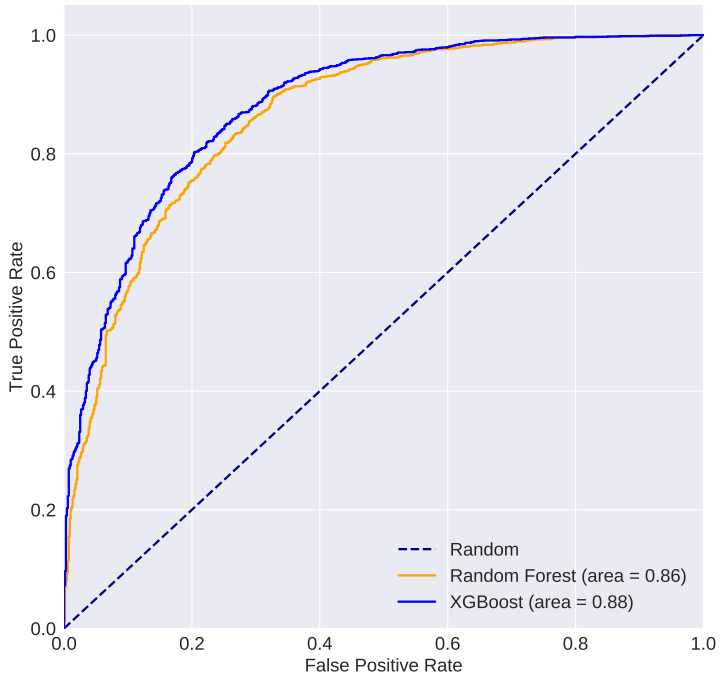


Unión de los datasets Integral y VarQ

Importancia transversal a todas las dimensiones estudiadas

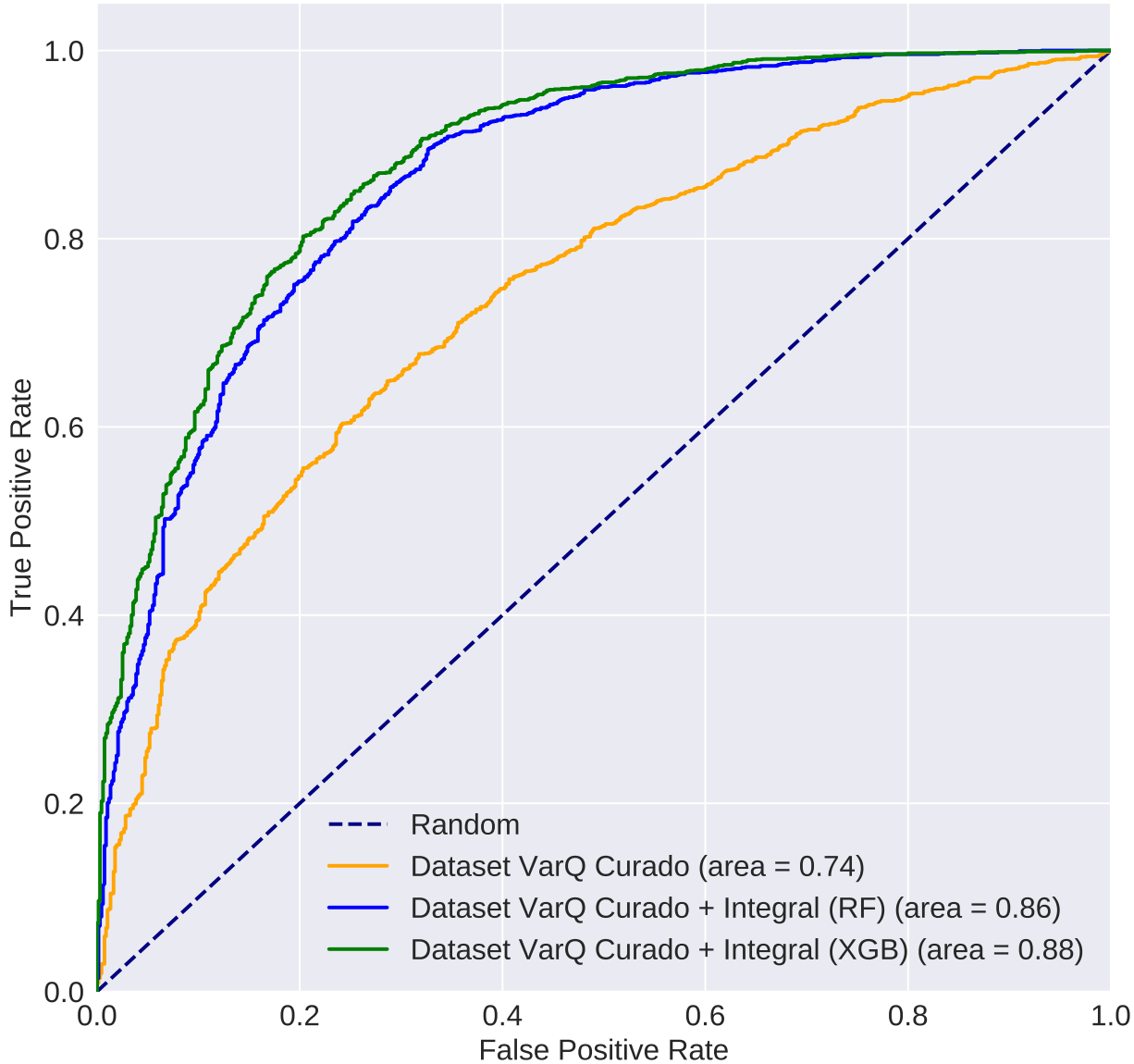


Importancia de features clusterizada usando rfpimp



Curva ROC

- **AUC VarQ+Integral: 0.88**
- AUC VarQ Curado: 0.74



Desempeño de los modelos usando las variantes de VarQ (Curado)

Conclusiones

- La combinación de distintas dimensiones del problema aportó excelentes resultados, consiguiendo un AUC de 0.90. Los features de conservación (genómicos y matrices de sustitución) fueron las que más aportaron al desempeño del modelo
- El método estándar de cálculo de importancia de features usado por scikit-learn puede ser engañoso en el caso de features altamente correlacionados
- La exploración de otros algoritmos más avanzados aportó mejoras sustanciales al modelo

Trabajo futuro

- Aumentar la cobertura de los features más importantes: La variación de la energía y las features de conservación genómica
- Mejorar la búsqueda de hiperparámetros en XGBoost
- Evaluar SNPs en regiones no codificantes (FATHMM-MKL)

¿Preguntas?

¡Muchas gracias!