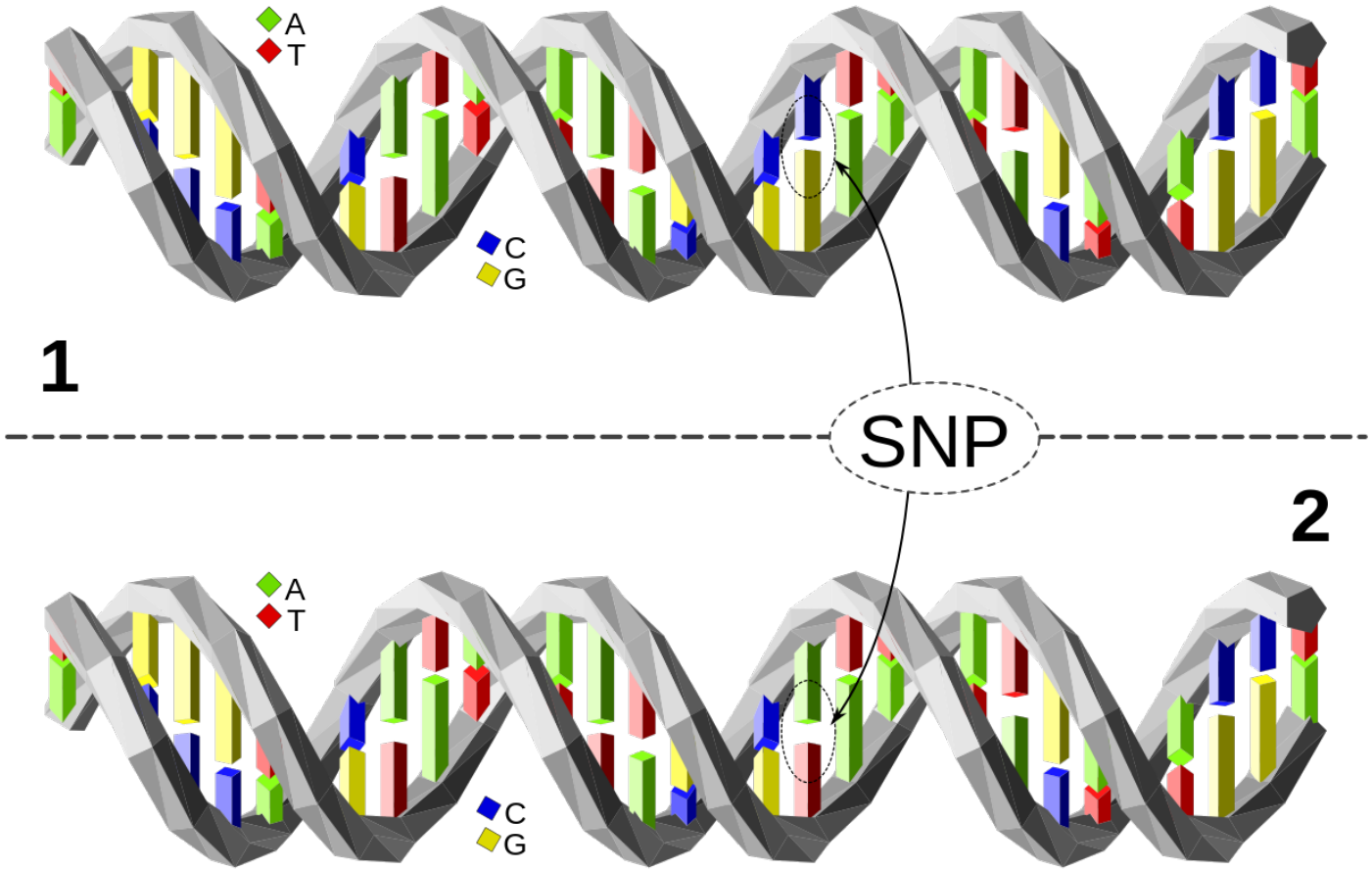
Predicción de patogenicidad en SNPs usando Aprendizaje Automático

Tesis de Licenciatura en Ciencias de la Computación

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

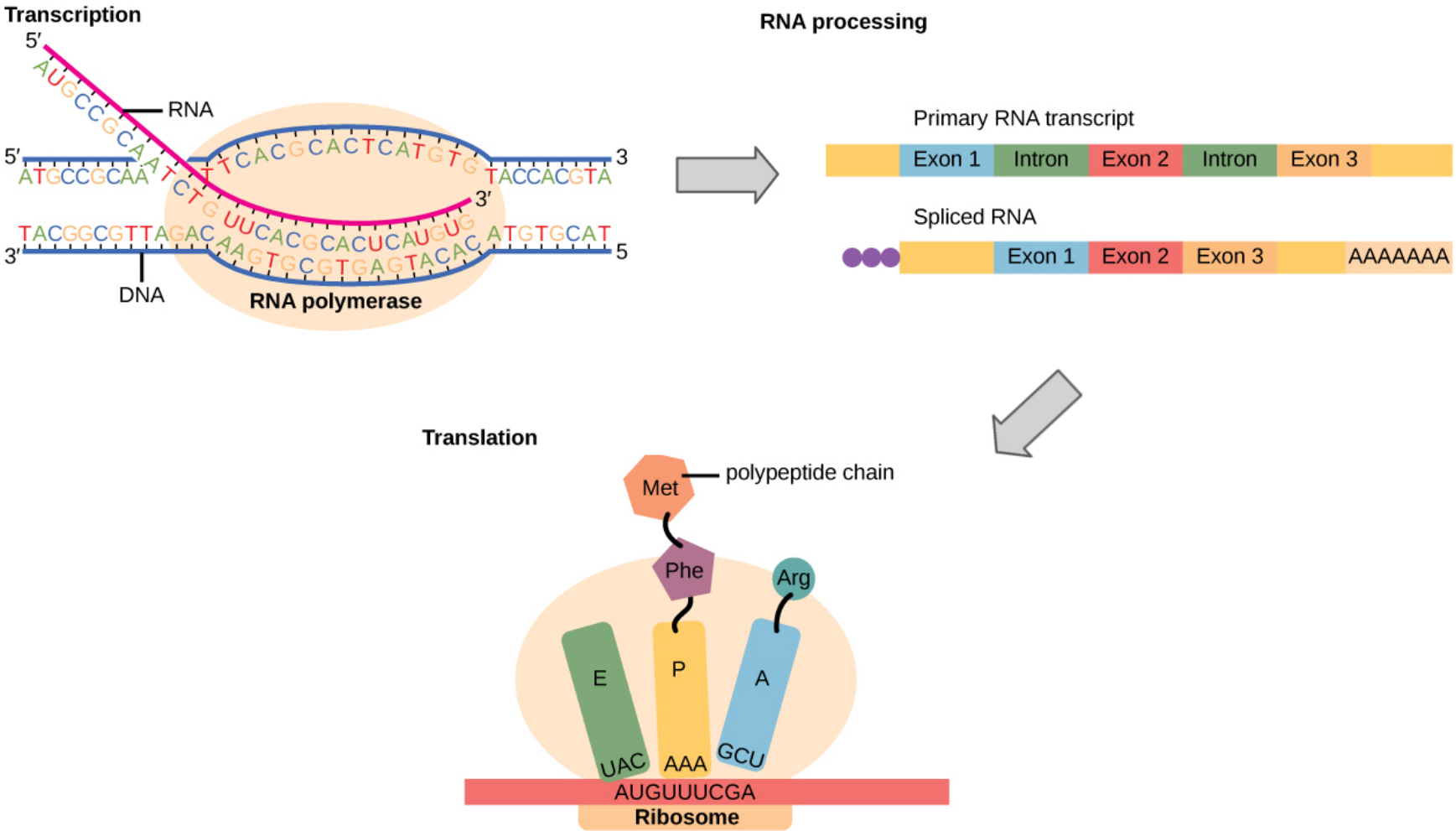
Departamento de Computación, FCEyN, UBA

Introducción: ¿Qué son los SNPs?



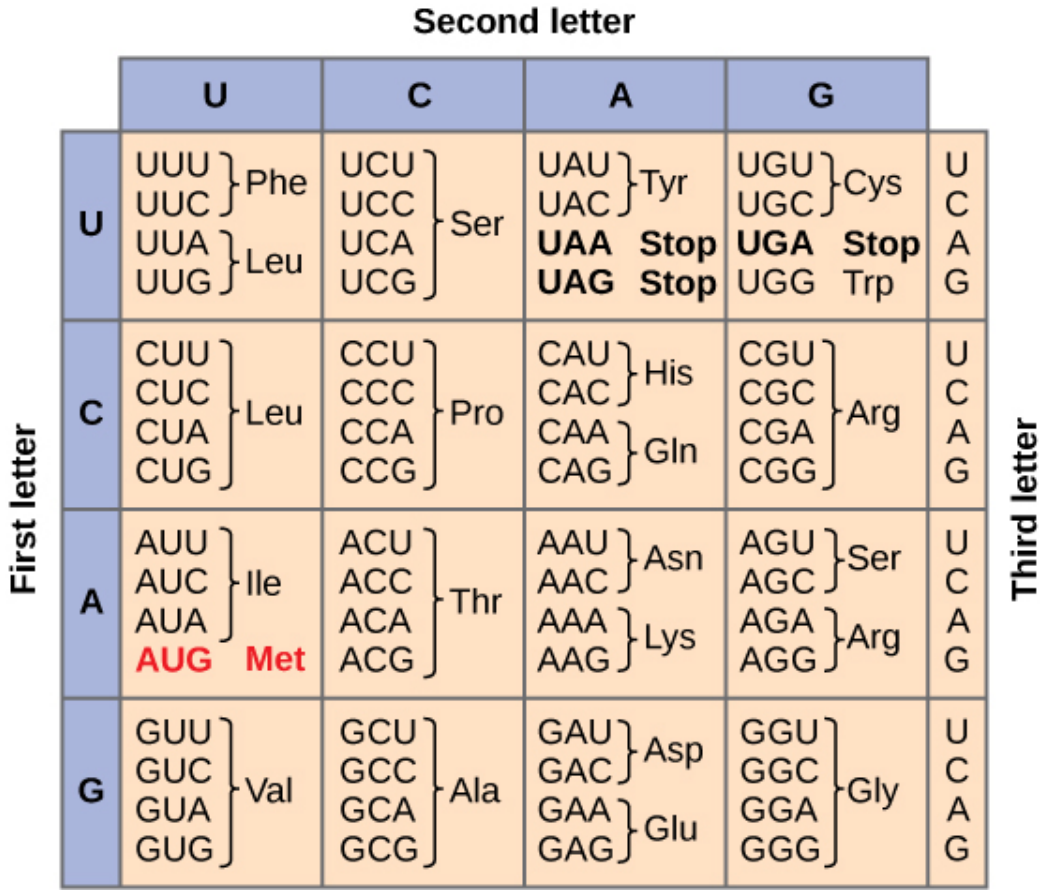
Single Nucleotide Polymorphism (SNP)

Introducción: Del ADN a las proteínas



Dogma central de la biología

¿Cómo se expresan los SNPs en el organismo?



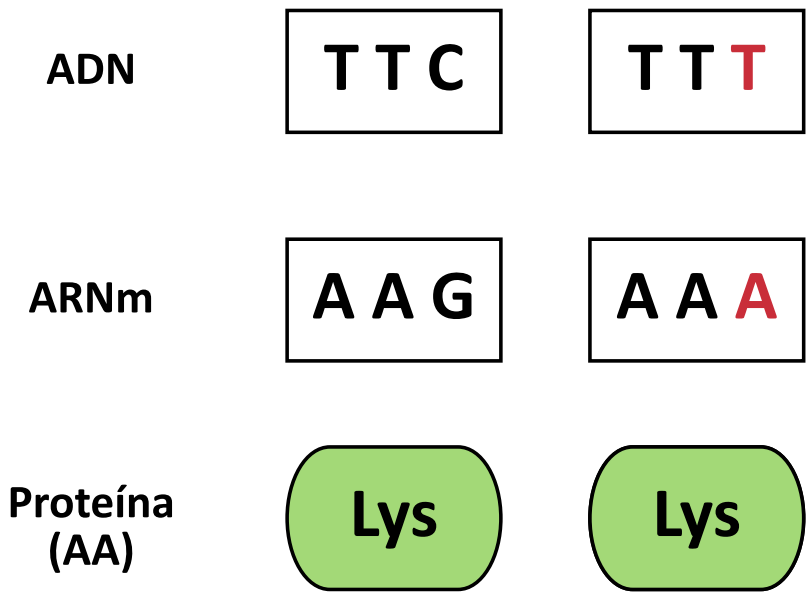
First letter \ Second letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr	UGU Cys UGC Cys	U C A G
	UUA Leu UUG Leu		**UAA Stop** **UAG Stop**	**UGA Stop** UGG Trp	
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln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

Tabla de codones de ARN

Introducción: Tipos de SNPs

Sustitución sinónima o *silent*

El cambio en el nucleótido no modifica el aminoácido

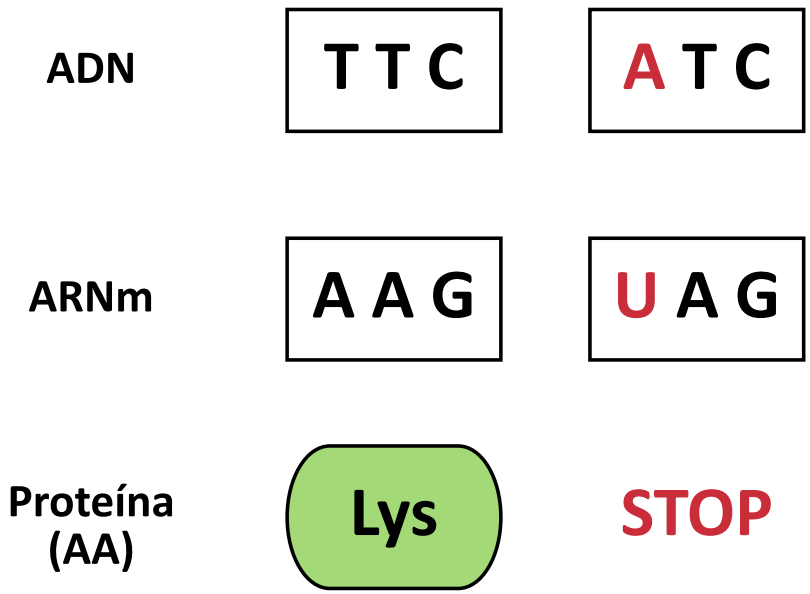


Sustitución *silent*

Introducción: Tipos de SNPs

Sustituciones no sinónimas

Nonsense: Generan un codón de terminación o *stop*

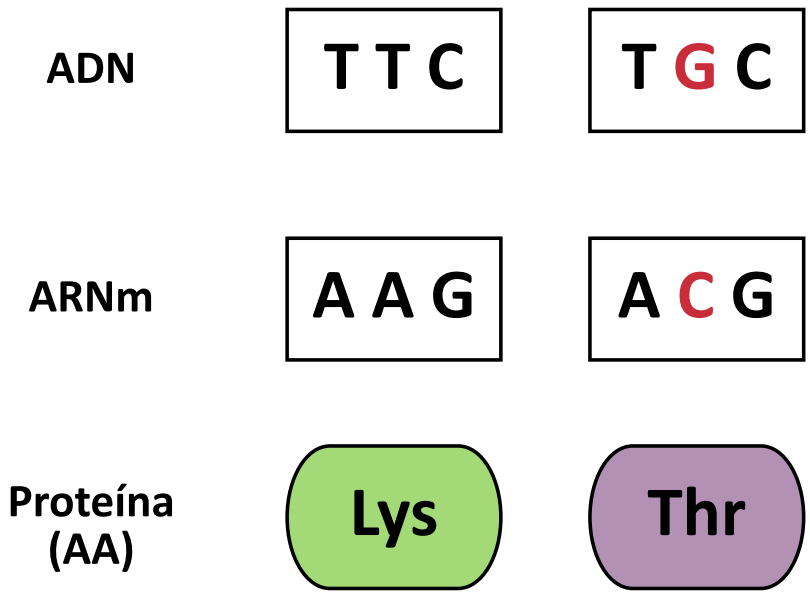


Sustitución *nonsense*

Introducción: Tipos de SNPs

Sustituciones no sinónimas

Missense: Generan un cambio de aminoácido en la proteína



Sustitución *missense*

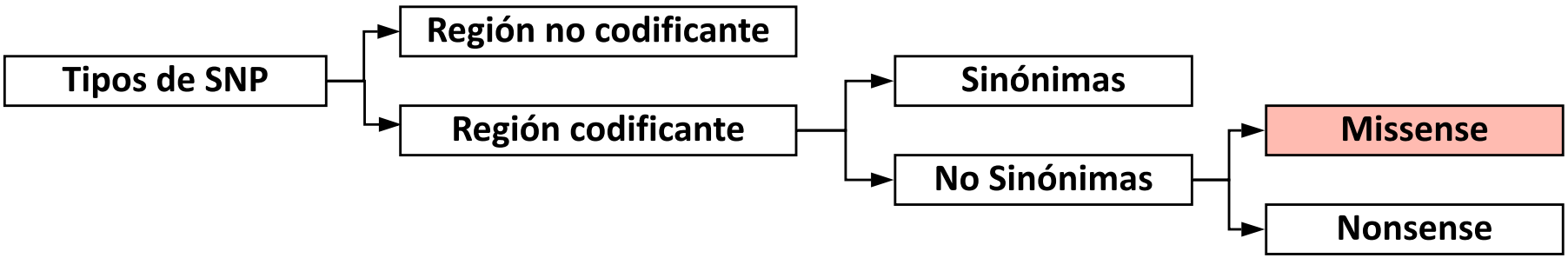
Foco de estudio: Variantes *missense*

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

Problema biológico: detectar la patogenicidad de SNPs

- La mayoría de las mutaciones no sinónimas son raras (AF $< .05\,\%$)
- Los estudios realizados con secuenciación tienen baja significación estadística
- Existen bases de datos biológicas que registran patogenicidad de mutaciones: Clinvar, Humsavar y otras

Main gene name	AA change	Type of variant	dbSNP
A4GALT	p.Pro251Leu	Polymorphism	rs28940571
A4GALT	p.Gln163Arg	Polymorphism	rs28915383
A4GNT	p.Ala218Asp	Polymorphism	rs2246945
AAAS	p.His160Arg	Disease	-
AAAS	p.Ser263Pro	Disease	rs121918550

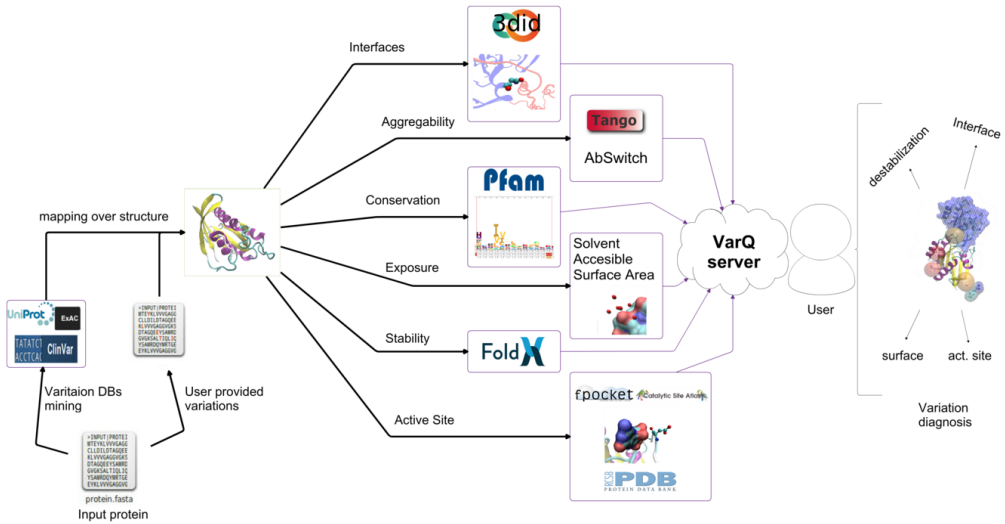
Selección de columnas de tabla Humsavar (extracto)

Enfoque computacional: un problema de clasificación

- **Objetivo: Predecir patogenicidad de SNPs *missense* humanos**
- Trabajos previos:
  - VEST (Carter et al., 2013)
  - FATHMM-MKL (Shihab et al., 2015)
  - REVEL (Ioannidis et al., 2016)
- Aprendizaje automático supervisado
- Dimensiones estructurales, físico-químicas de las proteínas, genómicas
- Análisis de importancia de las variables

¿Qué tan difícil es este problema?

Primer modelo: Propiedades estructurales usando VarQ



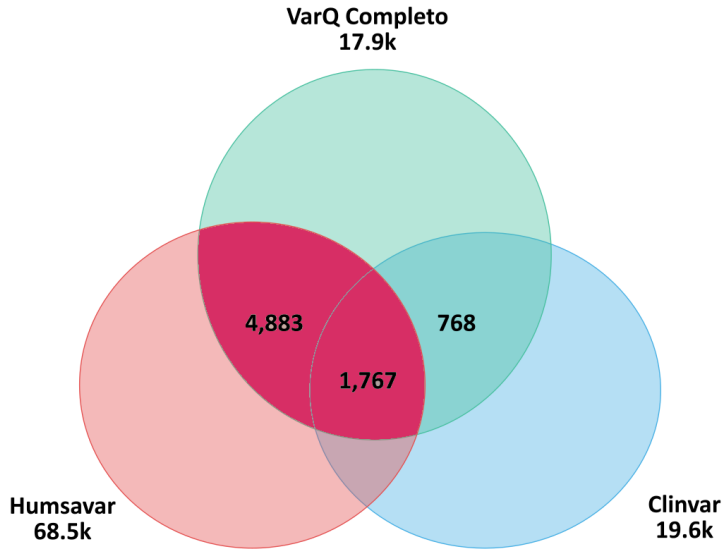
Pipeline de extracción de datos de VarQ

Features extraídos (cobertura)

- Variación de la energía
- SASA
- Porcentaje de SASA
- B-Factor
- Switchbility
- Aggregability
- Conservación
- Interfaz 3DID
- Interfaz PDB
- **Active Site**

Filtrado de variantes del dataset VarQ

- Removimos variantes sin un status confirmado (*risk factor*, *likely benign*, *uncertain significance*)
- Priorizamos con el reporte de Humsavar (Pathogenic, Disease)
- Aproximadamente 7,500 variantes: 72 % patogénicas, 28 % benignas



Intersección del dataset VarQ usando Humsavar y Clinvar

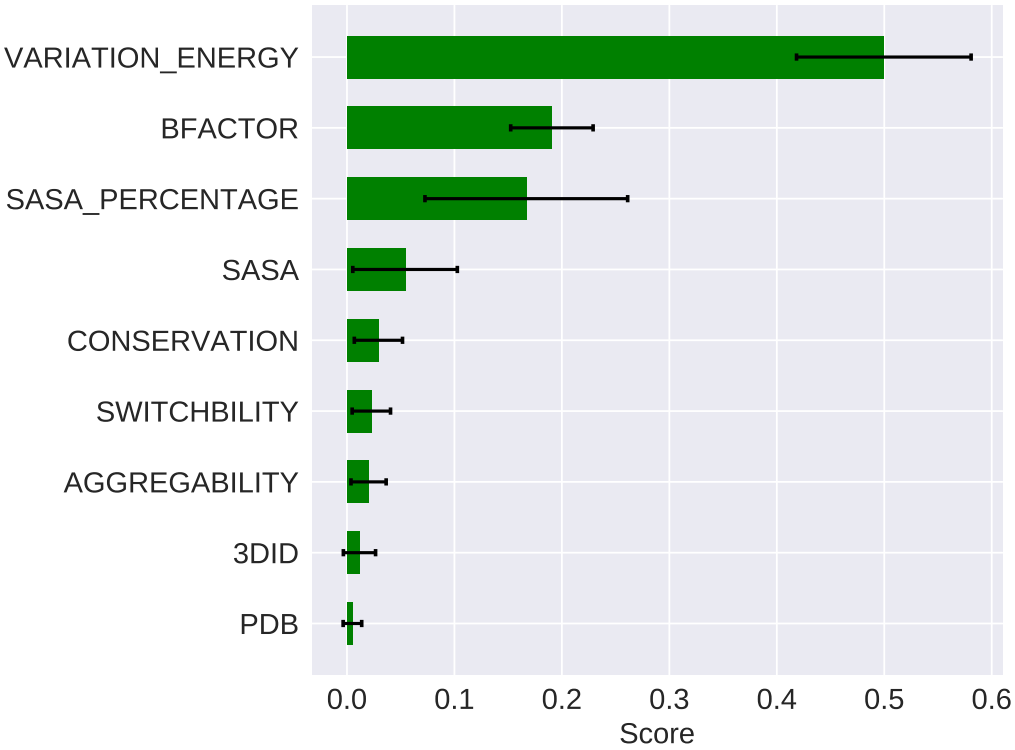
Generación de modelos de aprendizaje automático

- Modelos clásicos usando `scikit-learn`
  - Support Vector Classifier (kernel radial)
  - Random Forest
  - Regresión logística
- Imputación de variables nulas
- Búsqueda de hiperparámetros usando *3-fold Cross Validation*

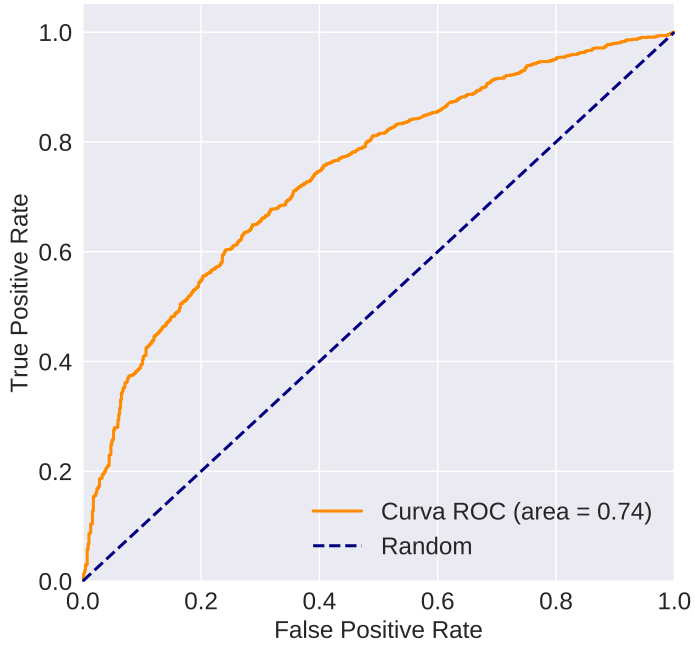
Comparación de modelos usando VarQ: Random Forest tiene el mejor AUC

	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

Resultados del modelo VarQ (Random Forest): La variable más importante es la Variación de la Energía



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

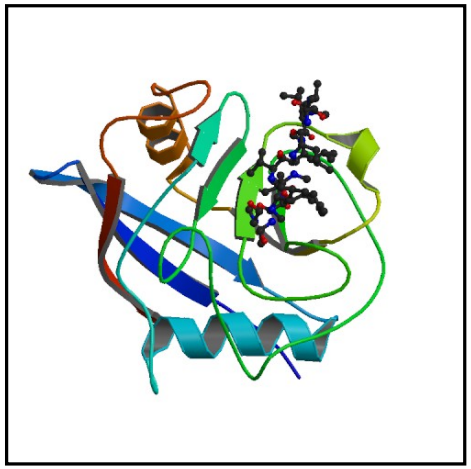


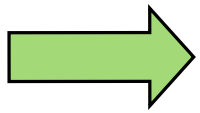
Curva ROC (0.74)

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

Modelo: Propiedades Físico-Químicas de la proteína

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





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

Extracción de secuencia proteica (ciclofilina) en formato FASTA usando Uniprot

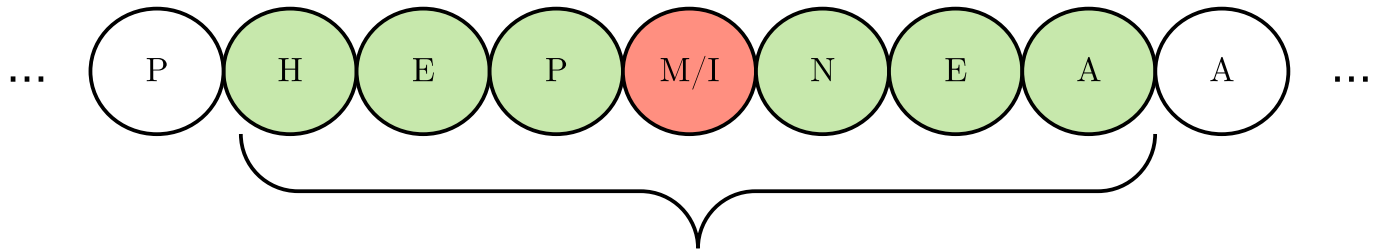
Generación de nuevas variables usando ProtParam

Parámetros calculados

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

Cambio en la variante

- Diferencia
- Log-ratio



Ventana de cálculo del parámetro

Variables físico-químicas extraídas de SNVBox

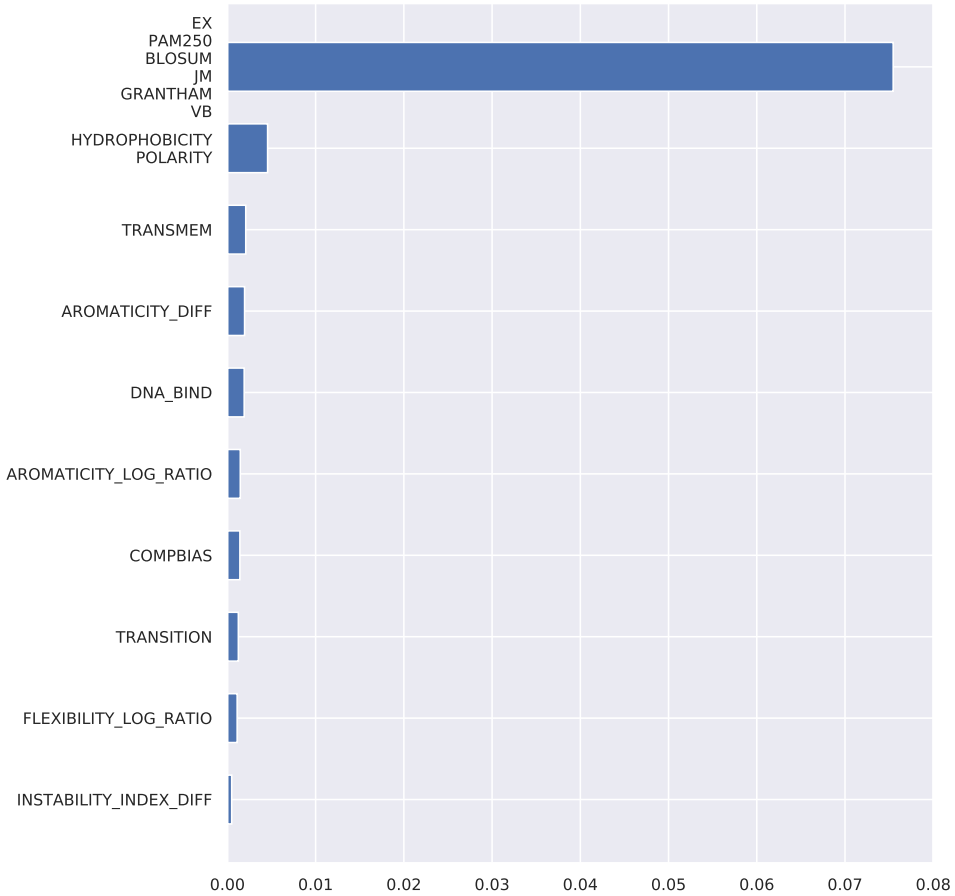
Variables a nivel de aminoácido (considerando sustitución)

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

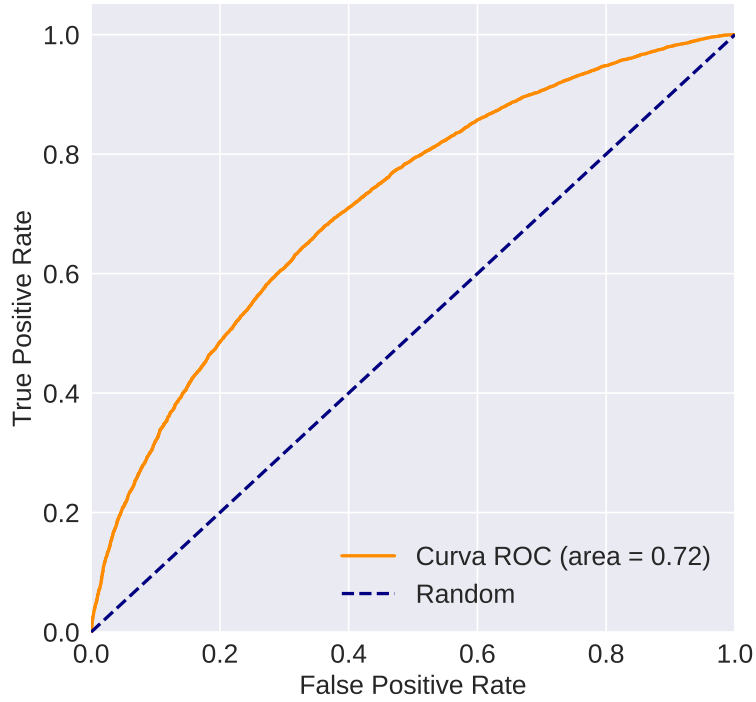
Variables a nivel de proteína (sin considerar sustitución)

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

Las matrices fueron las más relevantes



Importancia de variables clusterizada usando `rfpimp`



Curva ROC (0.72)

¿Cuál es el valor predictivo de las variables genómicas?

Modelo: Variables genómicas

- Identificador rsID: aproximadamente 55,000 variantes en Humsavar
  - 68 % variantes benignas
  - 32 % variantes patogénicas
- Fuentes de variables:
  - SNVBox
  - dbSNP
  - Genome Browser (UCSC)

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

Variables del modelo Genómico

Variables de conservación genómica

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

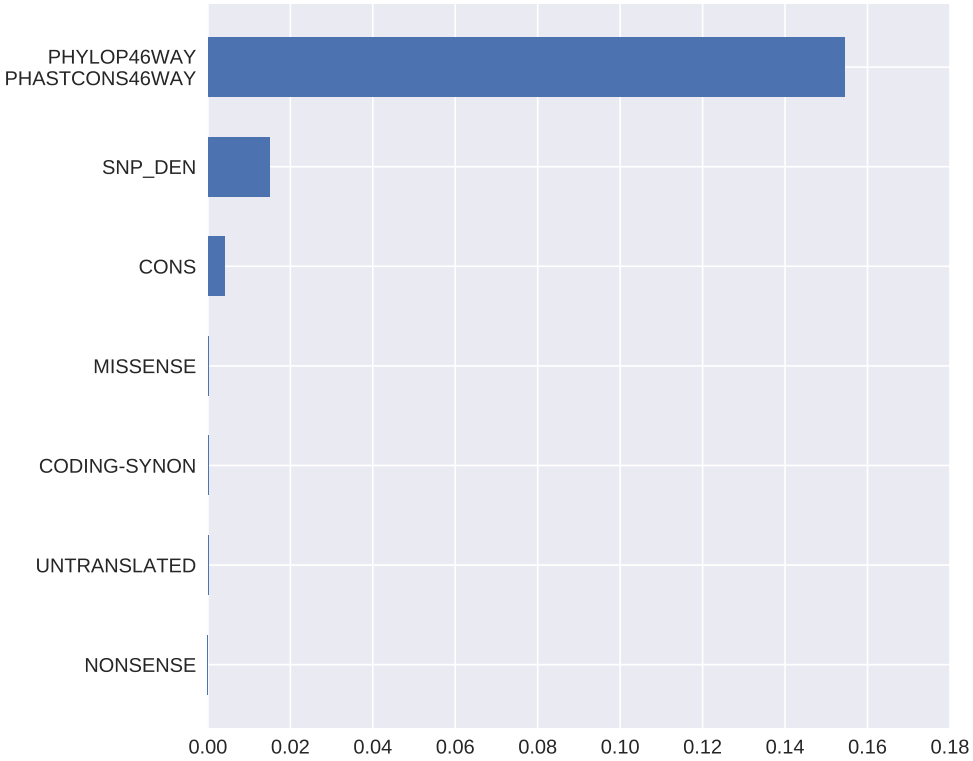
Variables extraídas de SNVBox

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

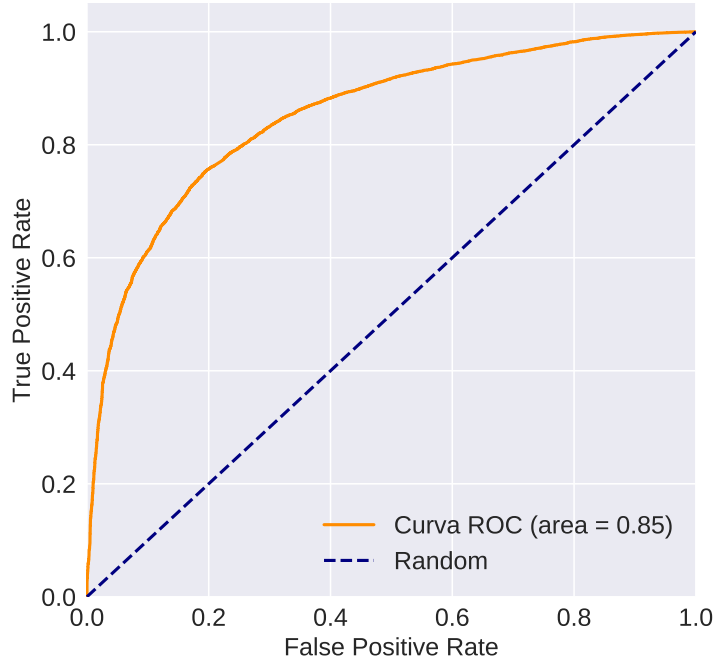
Variables relativas a la clase funcional

- Missense
- Nonsense
- Intrón
- y otras

La conservación genómica es importantísima!



Importancia de variables clusterizada usando `rfpimp`

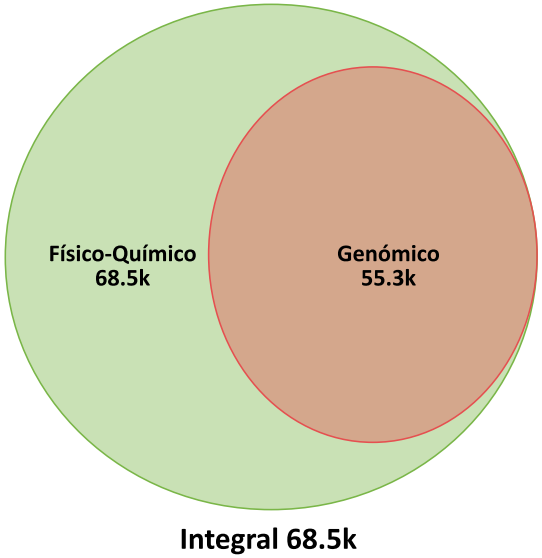


Curva ROC (**0.85!**)

Podemos mejorar el modelo genómico integrando las variables físico-químicas?

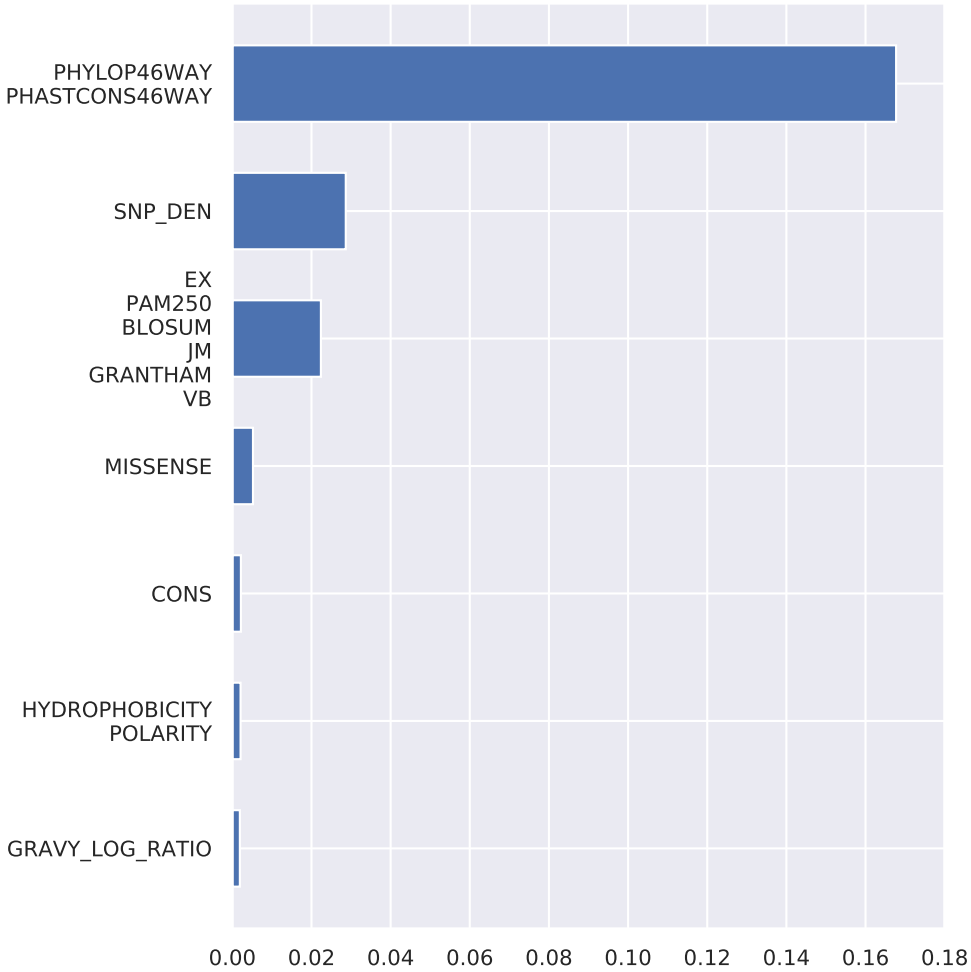
Integrando las variables físico-químicas y genómicas

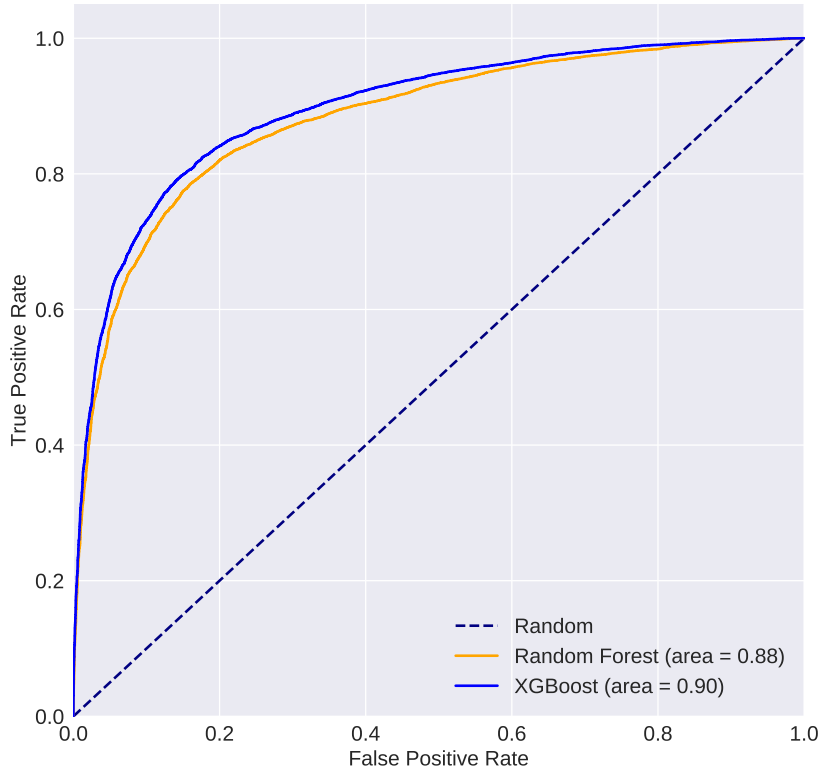
- Dataset Humsavar: 68 mil variantes
- Cobertura variables genómicas: aprox. 80 %
- Cobertura variable físico-químicas: misma que el dataset físico-químico
- Evaluamos un nuevo método de aprendizaje automático: XGBoost



Unión de los datasets Físico-Químico y Genómico

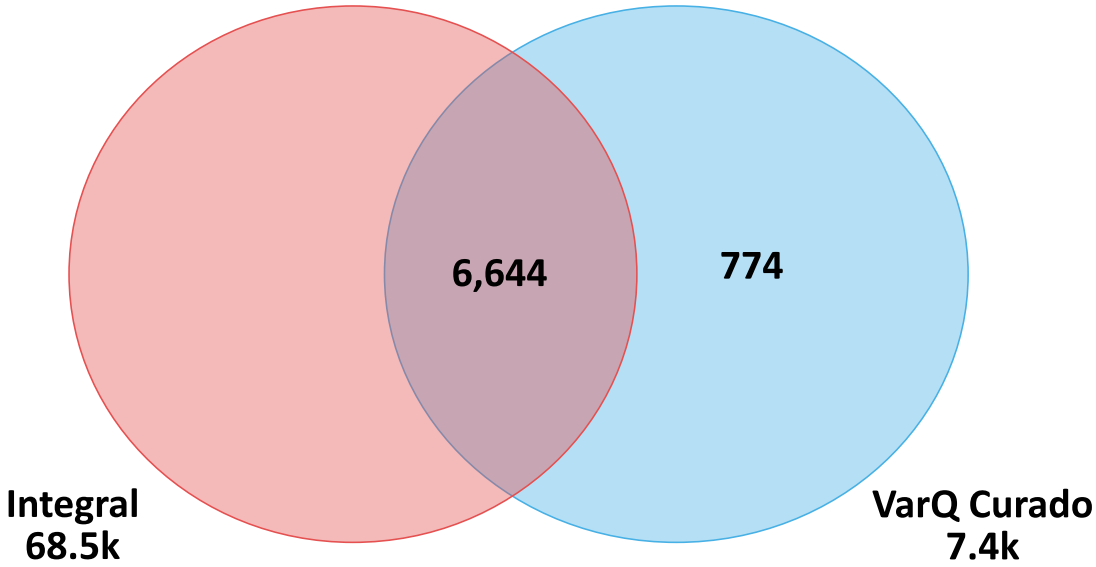
XGBoost supera a Random Forest





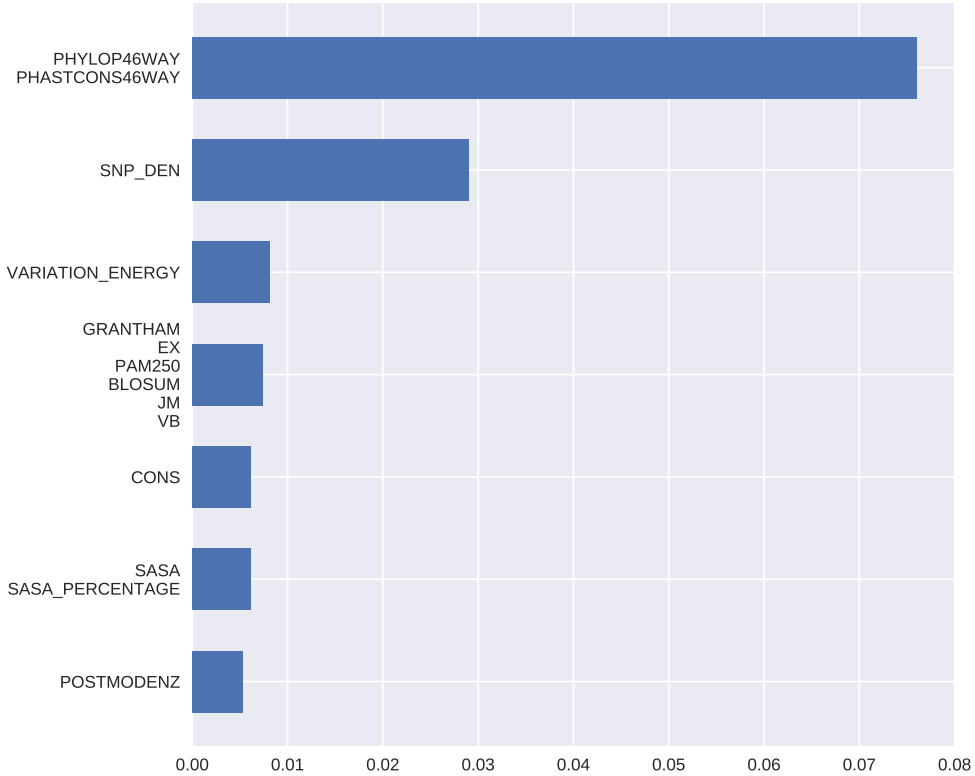
AUC: 0.90

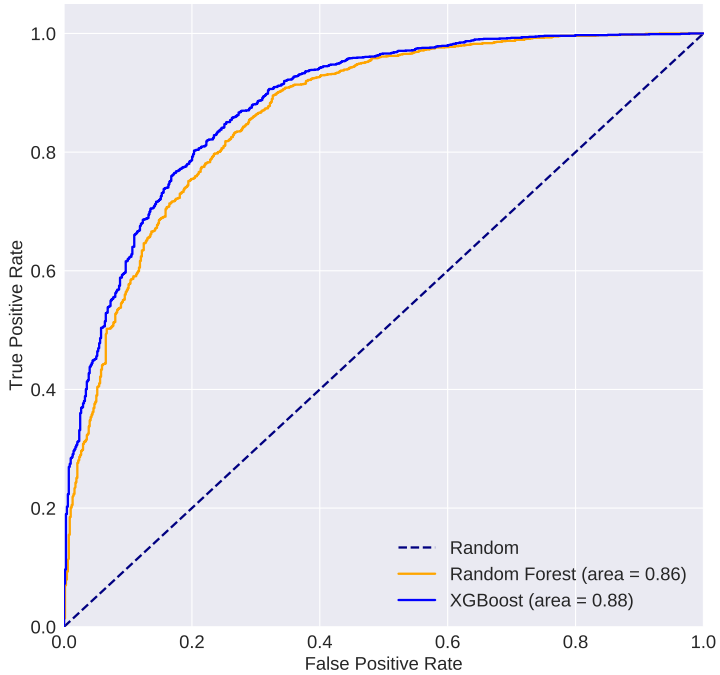
Modelo Integral + VarQ



Unión de los datasets Integral y VarQ

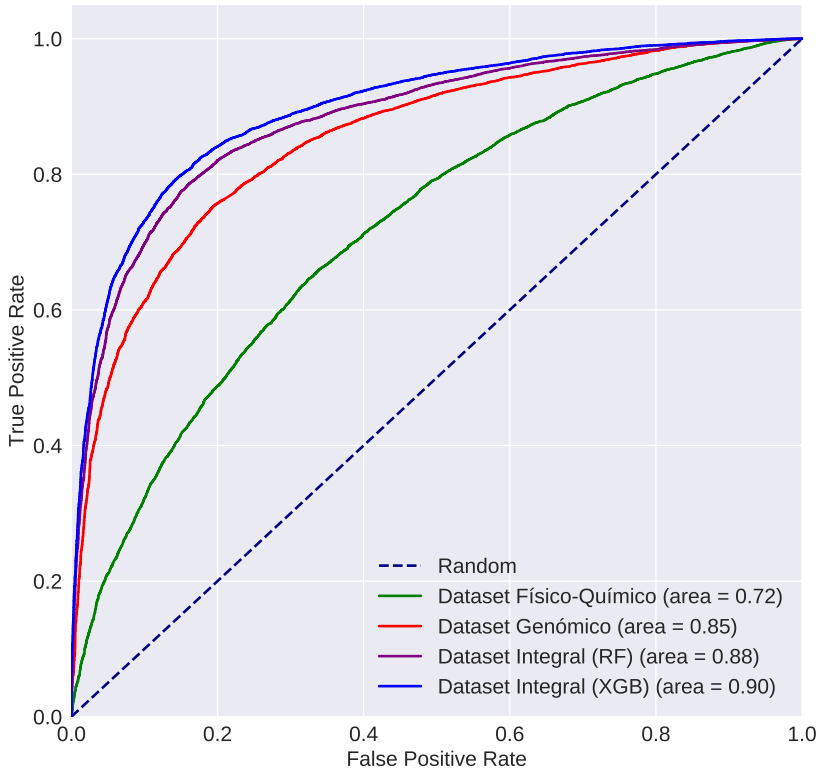
Resultados del modelo Integral + VarQ (XGBoost)



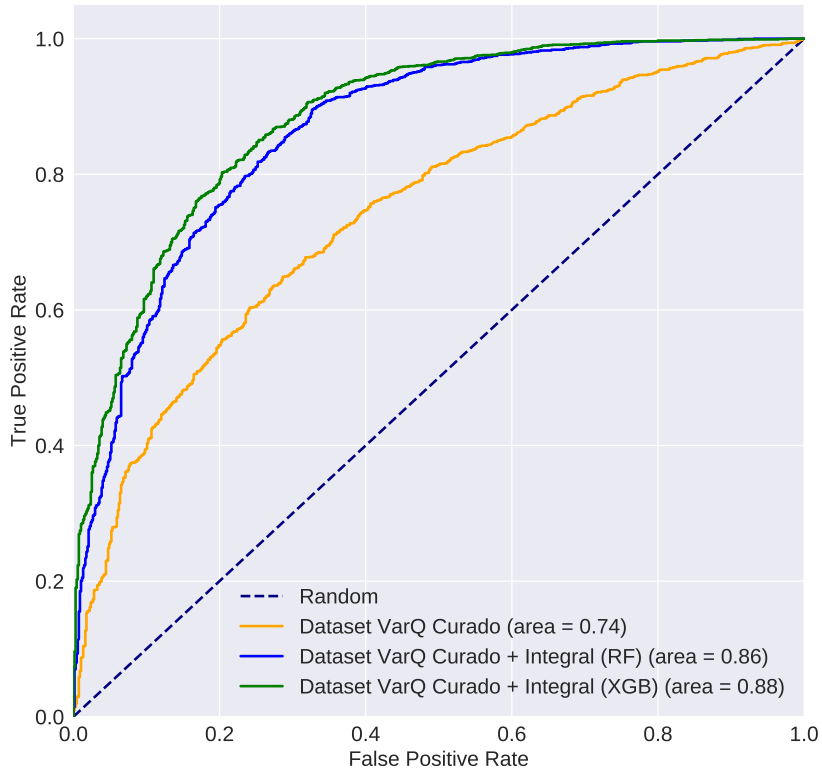


AUC: 0.88

Comparación entre los distintos modelos



Dataset Humsavar



Dataset VarQ (Curado)

Conclusiones:

- La combinación de distintas dimensiones del problema aportó buenos resultados, consiguiendo un AUC de 0.90
- El método estándar de cálculo de importancia de variables usado por scikit-learn puede ser engañoso en el caso de variables altamente correlacionadas
- Los mejores resultados fueron obtenidos por algoritmos de Boosting

Trabajo futuro

- Aumentar la cobertura de las variables más importantes: La variación de la energía y las variables de conservación genómica
- Mejorar la búsqueda de hiperparámetros en XGBoost
- Evaluar SNPs *nonsense* o no codificantes
- Mejoras metodológicas

¿Preguntas?

¡Muchas gracias!