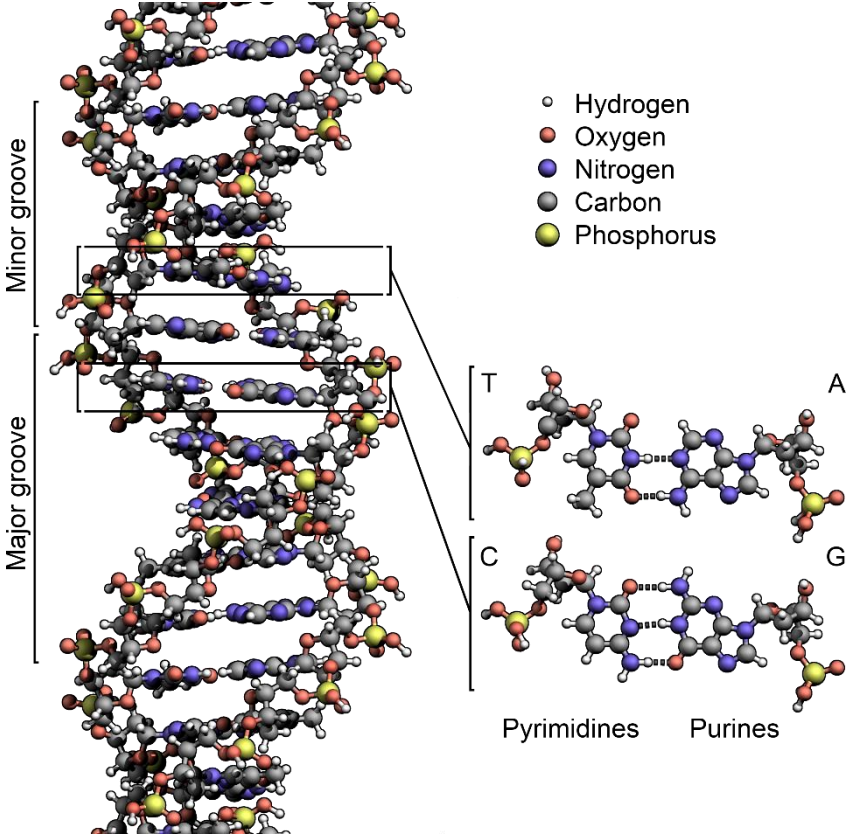
Predicting pathogenesis of missense mutations in membrane proteins

ESCI Resaarch Seminar - 09/06/2021

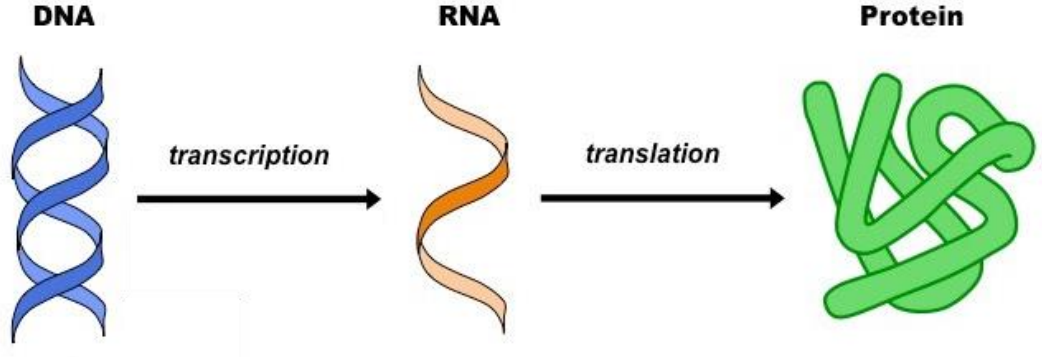
Arnau Cordomí
arnau.cordomi@esci.upf.edu



https://commons.wikimedia.org/wiki/File:DNA_Structure%2BKey%2BLabelled.pn_NoBB.png

Humane genome: 3,100,000,000 base pairs
(diploid 2x) 6,200,000,000 base paris

DNA to RNA to Protein



Store information

FUNCTION

AGCTACGTACGTA
CC**GGTAGA**CGAG
GATCGACTAGTCG
ATAGTAGCTACG

GGUAGA

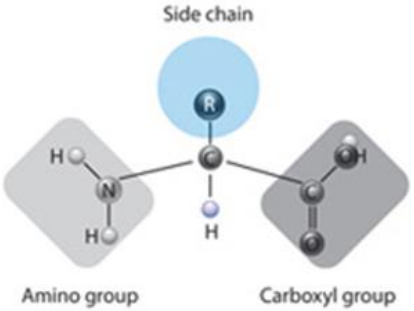
GR

A, C, G, T

A, C, G, U

A, C, D, E, F, G, H, I, K, L
M, N, P, Q, R, S, TW, Y, V

Building blocks of proteins



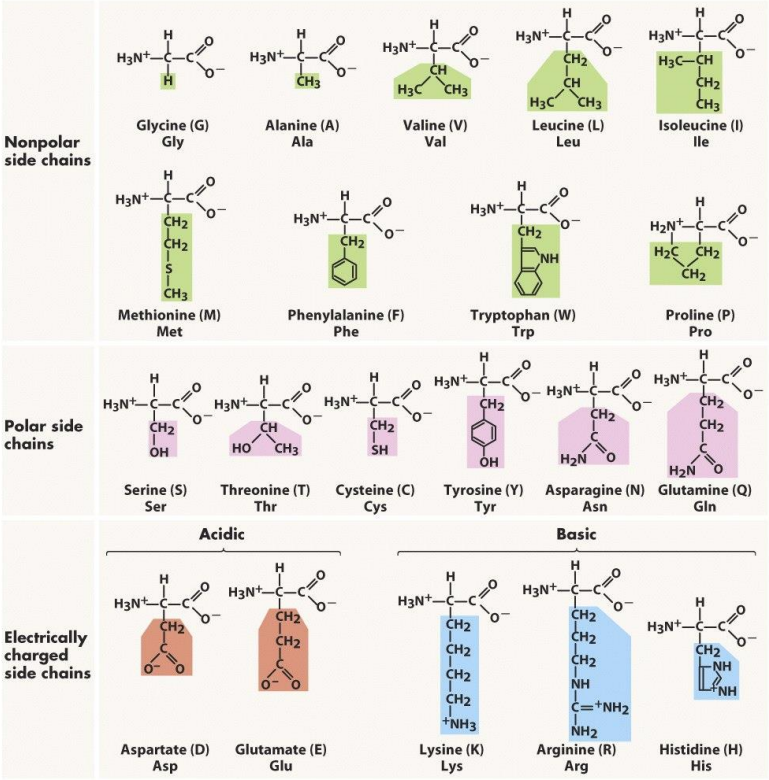
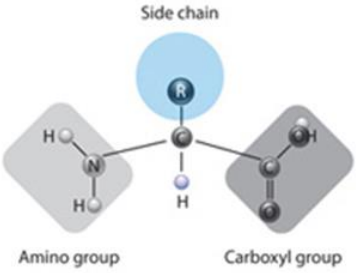
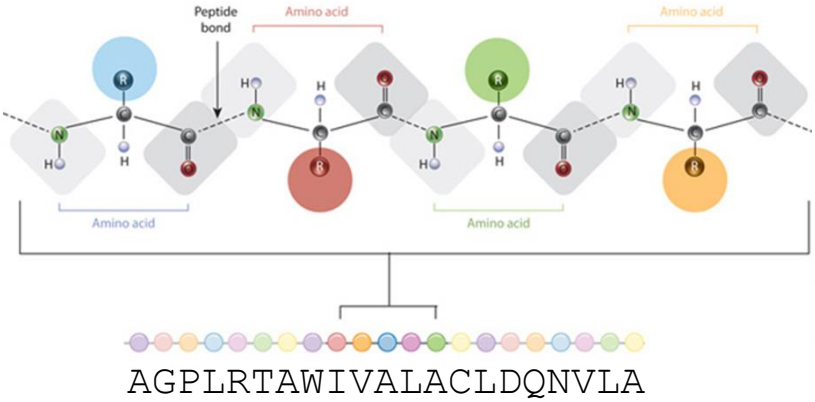
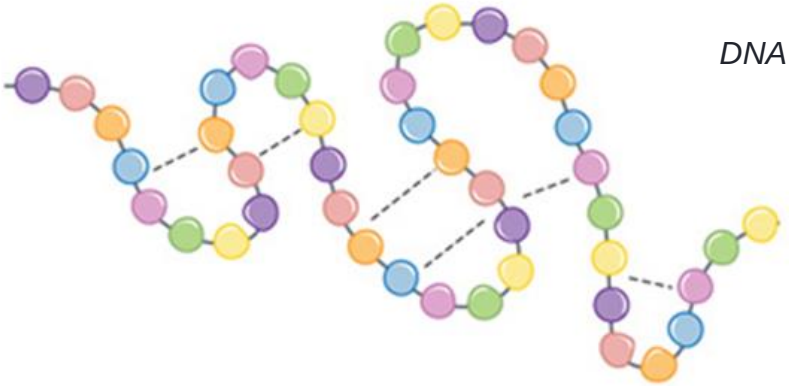


Figure 3-5 Biological Science, 2/e

Tertiary structure





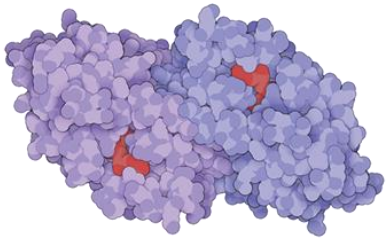


DNA sequence of the gene determines the amino acid sequence

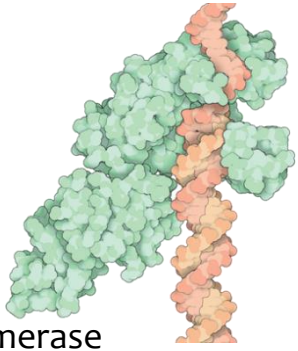
amino acid sequence determines protein 3D structure

Proteins are molecular machines

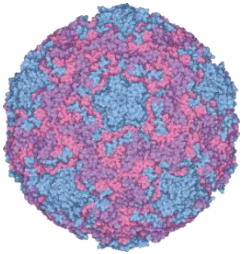
https://cdn.rcsb.org/pdb101/molecular-machinery/



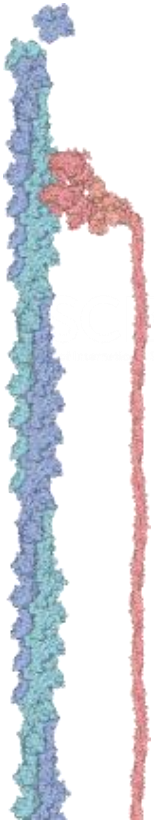
Alcohol dehydrogenase



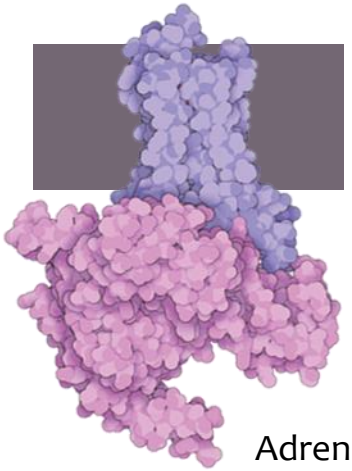
DNA polymerase



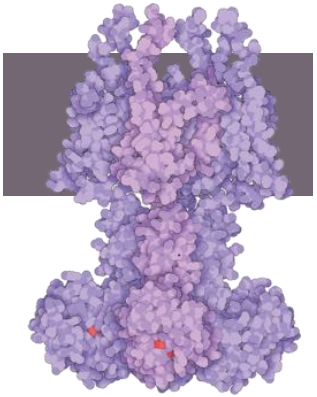
Virus capsid



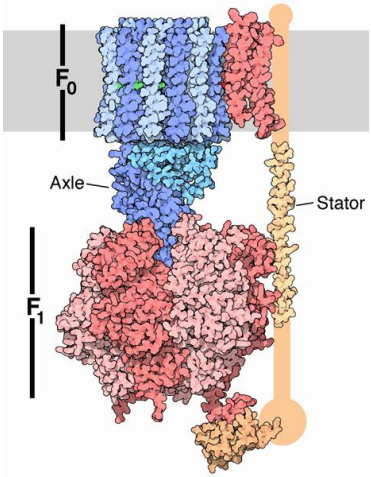
Actin and myosin



Adrenergic receptor

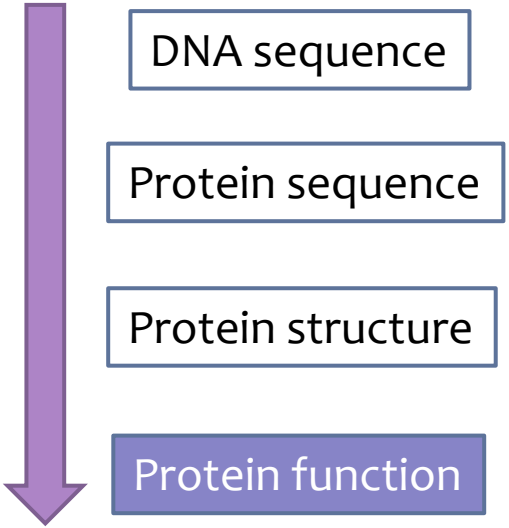


Potassium channel



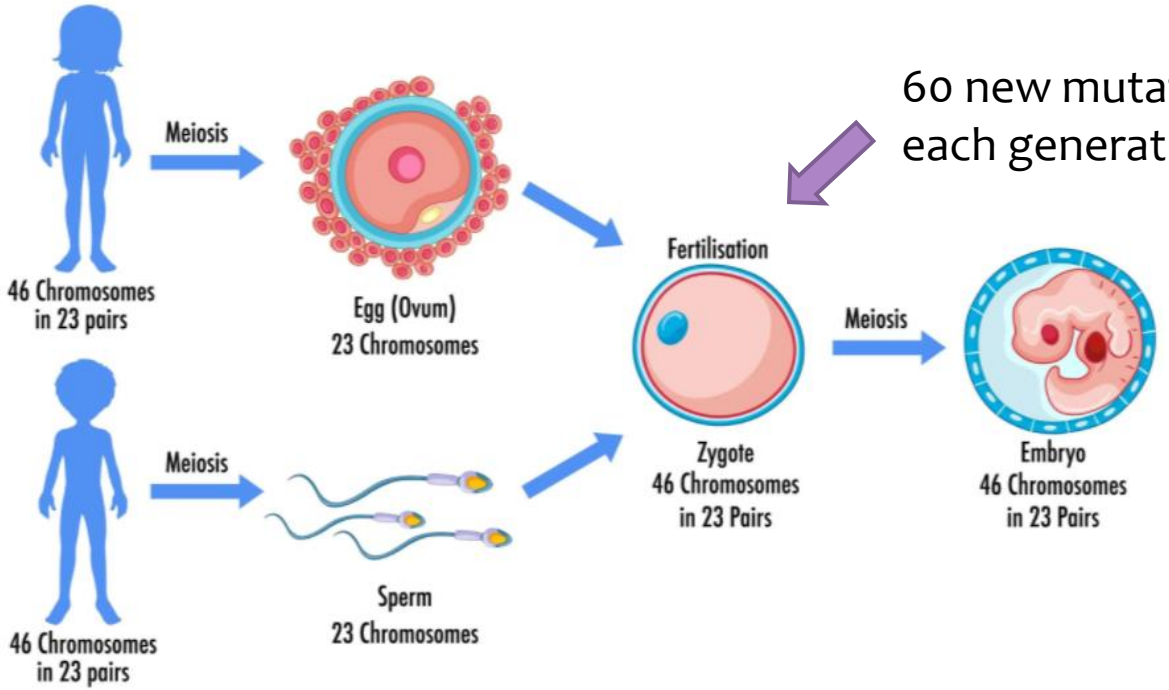
ATP Synthase

DNA sequence determines functoin

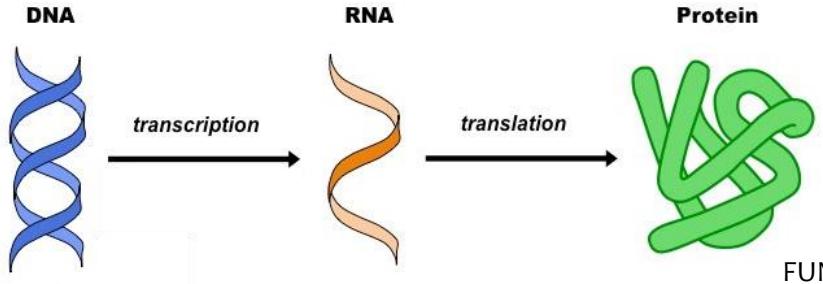


Human genetic variation

Human genome	3,100,000,000 base pairs
Typical difference between individuals	20,000,000 base pairs (0.6%)
Genetic variation:	> 300,000,000 variants

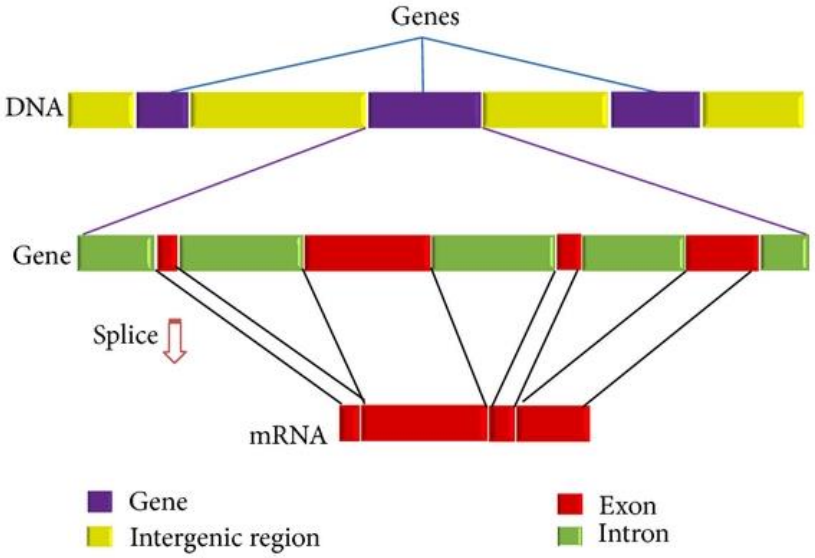

60 new mutations each generation
Meiosis
46 Chromosomes in 23 pairs
Egg (Ovum) 23 Chromosomes
Fertilisation
Zygote 46 Chromosomes in 23 Pairs
Meiosis
Embryo 46 Chromosomes in 23 Pairs
Meiosis
46 Chromosomes in 23 pairs
Sperm 23 Chromosomes

Will mutations have a functional effect?



Store information

FUNCTION



The genetic code is degenerated

GCTAGCTATCTA**TAT**GGGATATGA **UAU : Y**

First base position	Second base position: U		Second base position: C		Second base position: A		Second base position: G		Third base position
U	UUU	[1]P	UCU	S	UAU	Y	UGU	C	U
	UUC		UCC		UAC		UGC		C
	UUA	L	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	W	G
C	CUU	L	CCU	P	CAU	H	CGU	R	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	I	ACU	T	AAU	N	AGU	S	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	K	AGA	R	A
	AUG	M	ACG		AAG		AGG		G
G	GUU	V	GCU	A	GAU	D	GGU	G	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	E	GGA		A
	GUG		GCG		GAG		GGG		G

[1]The one letter symbol of amino acids.

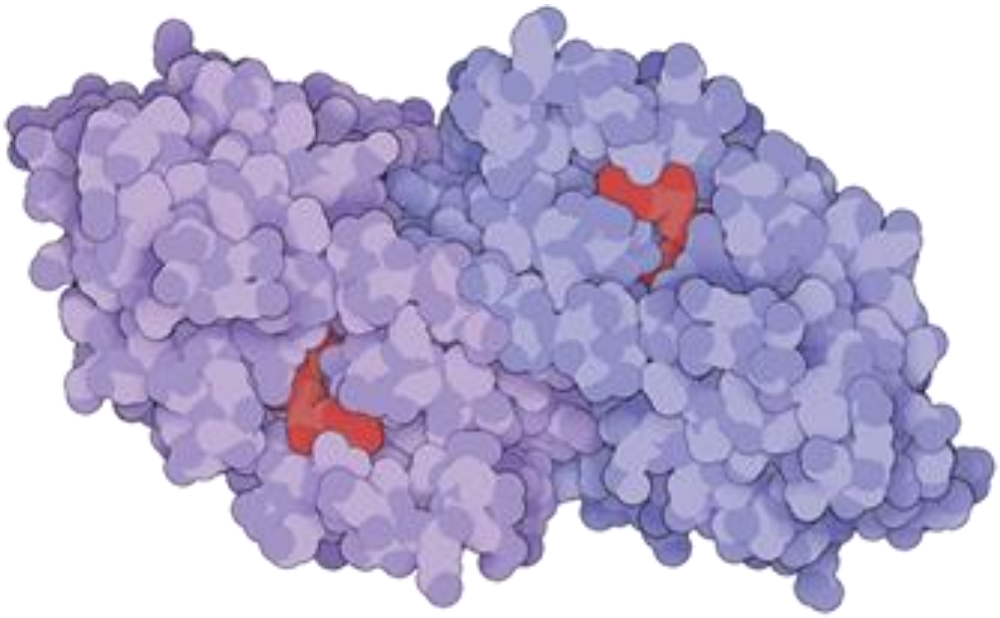
UA**U** → UA**C** : Y

Synonymous

UA**U** → U**G**U : C

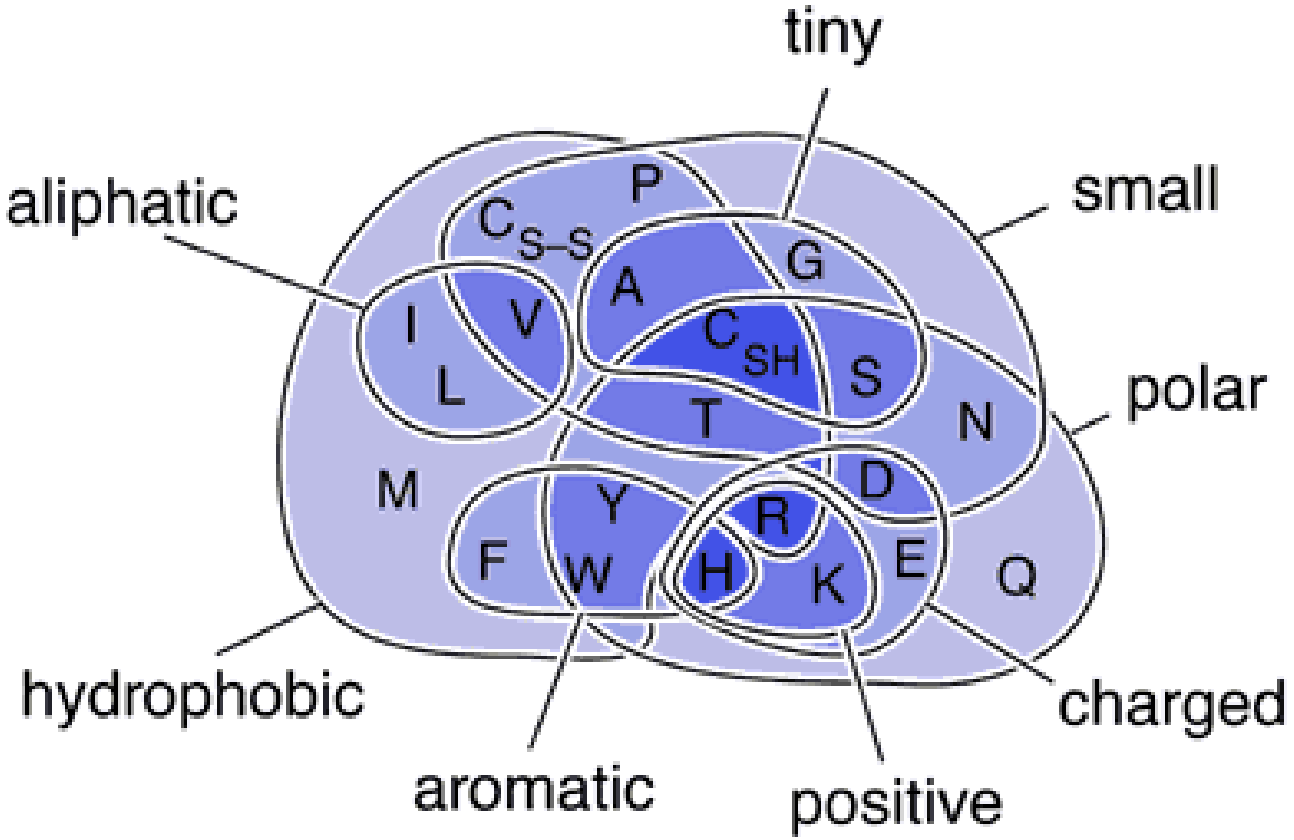
Missense

Will the mutation have any effect?

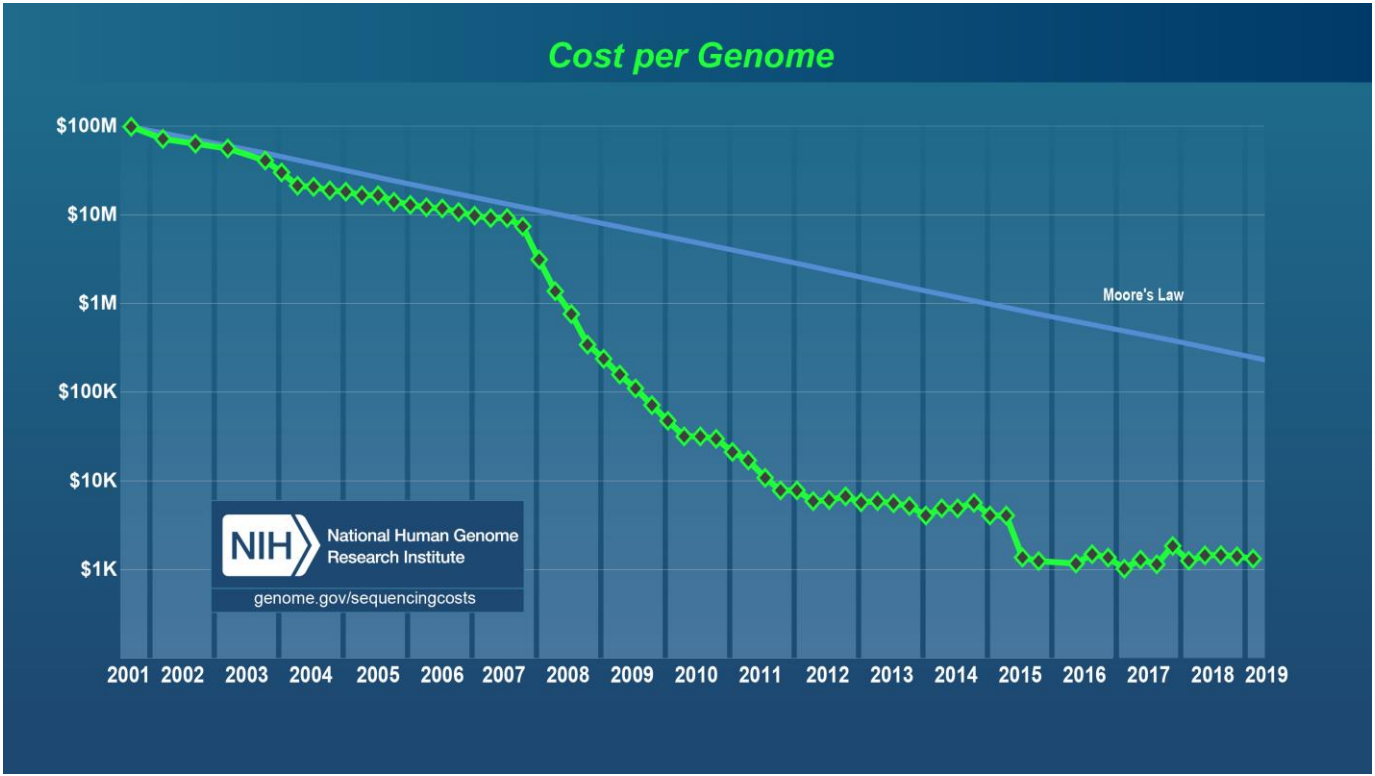


Alcohol dehydrogenase

Amino acids grouped by physicochemical properties



https://commons.wikimedia.org/wiki/File:Amino_Acids_Venn_Diagram.png



genome sequencing and analysis to reveal the genetic basis of disease in patients

Mardis *Genome Medicine* 2010, 2:84
http://genomemedicine.com/content/2/11/84

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

20,000,000 base pairs (0.6%)

Where is the mutation causing the disease?

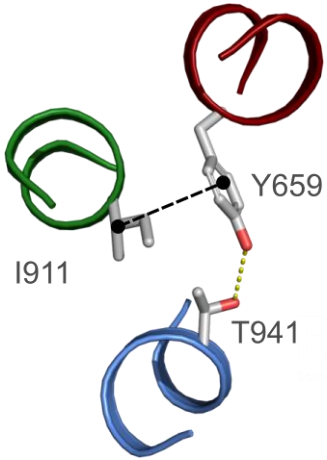
Comparison

DNA sequence

Protein sequence

Protein structure

Protein function



Comparing sequences is very cheap!

KNOWN QMKEDAKGKSEEELAECFRIFDRNMDGYIDAEELAEI

UNKNOWN QMKEDAKGKSEEELAECFRIFDRNMDG**F**IDAEELAEI

Multiple sequence alignment

Compare related proteins

```
CMKTDSKGKSEDELSDLFRMFDKNADGYIDAEELADL
QMKEDAKGKSEEELAECFRIFDRNMDGYIDAEELAEI
QMKADAKGKSEEELAECFRMFDKNADGYIDLDELADV
QMKSDAKGKSEEELAECFRIFDRNMDGYIDAEELAEI
```

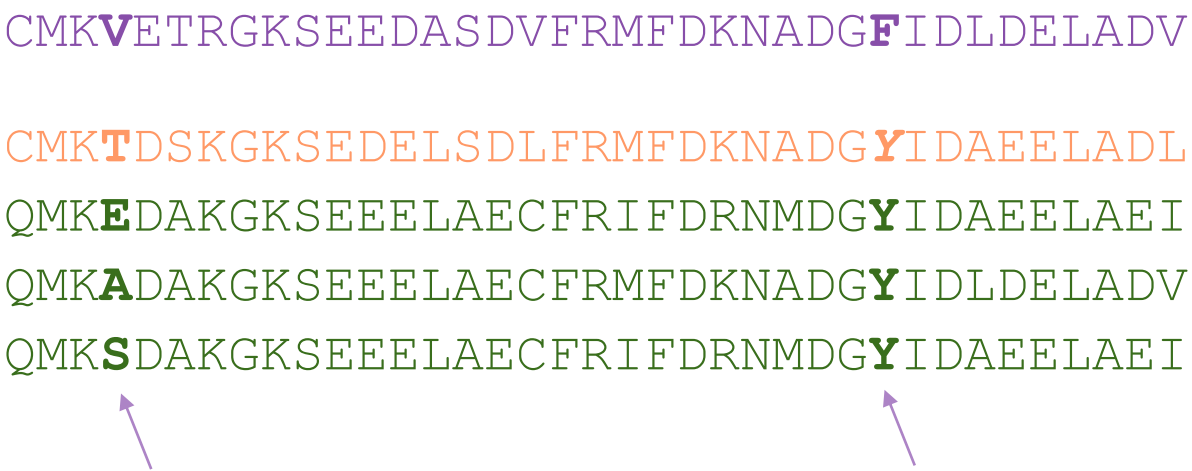
Not important?

Y required

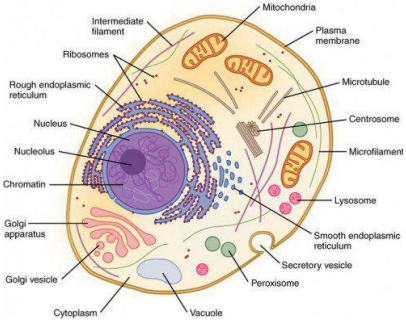
Each line : a protein
Each column: the equivalent position in each protein

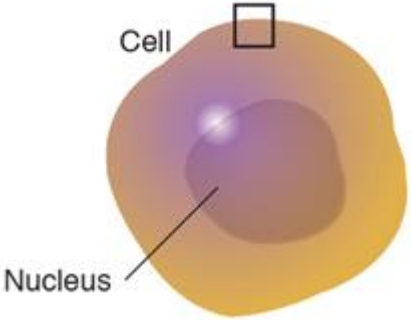
A protein family

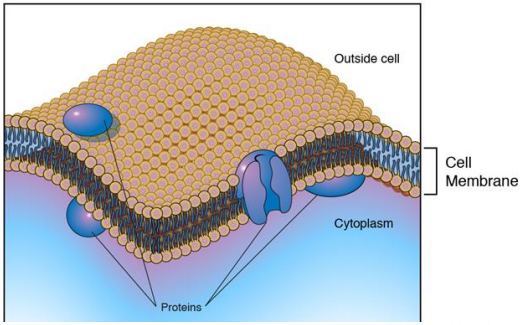
Multiple sequence alignment

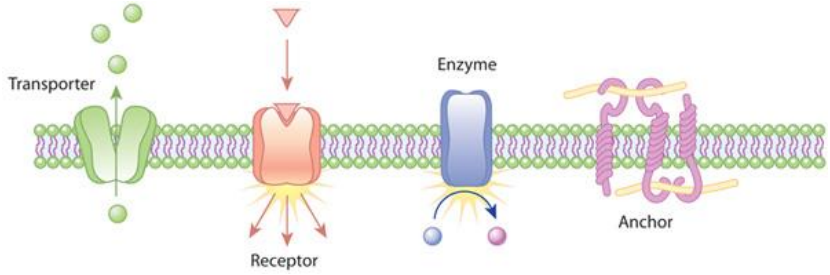


Focus on membrane proteins









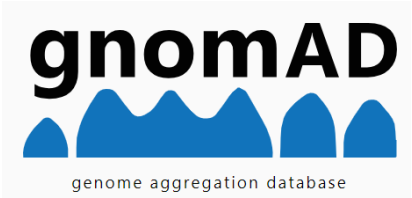
- 25% of proteins in the human genome are membrane proteins
- 50% of drugs target membrane proteins

GOAL: Predict pathogenesis of mutations in membrane proteins (only the regions that cross the membrane)

Input data

Protein database

List of all membrane proteins



Genome database (140,000 individuals

Database of variants with clinical information (500,000 variants)

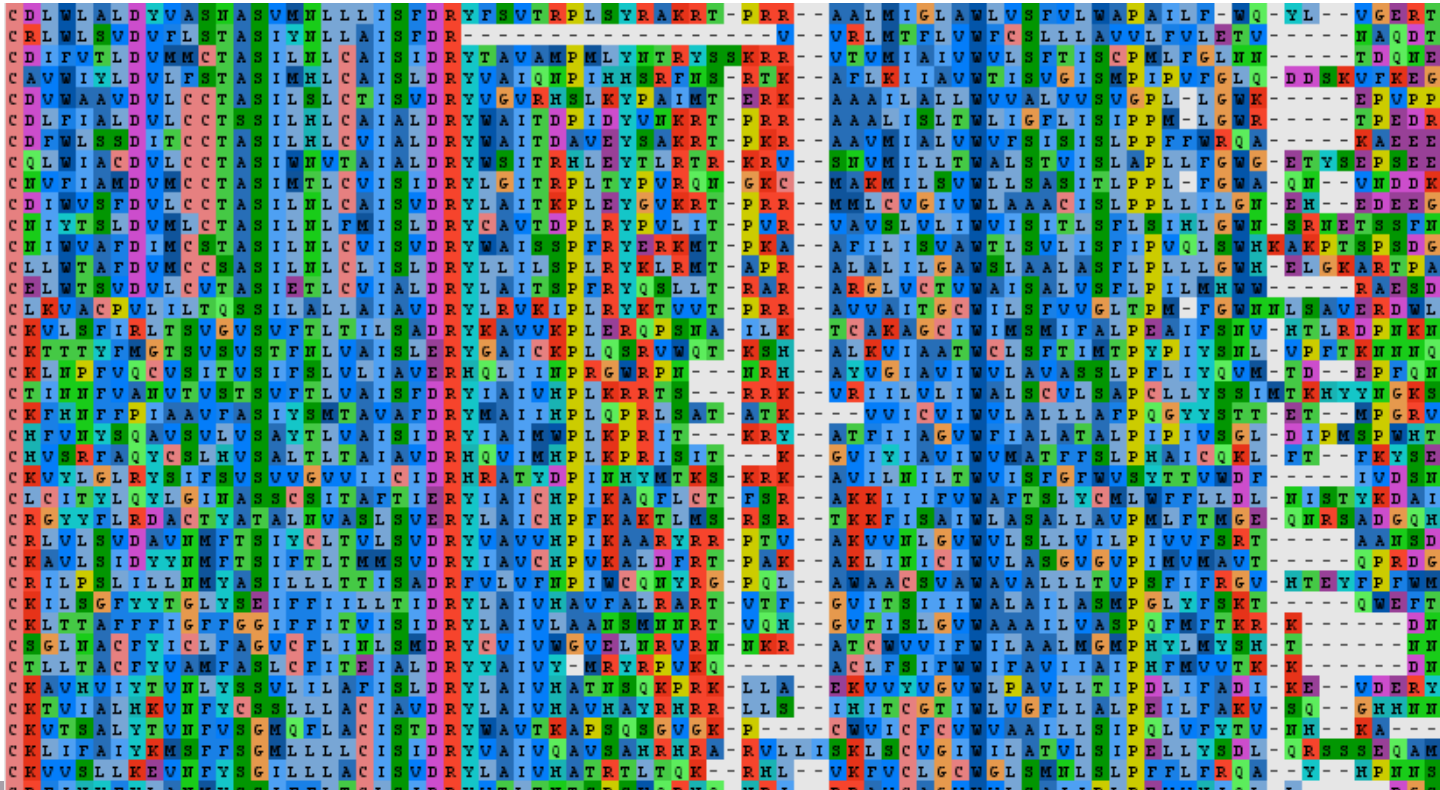
List of all mutations with labels (pathogenic or non pathogenic

Pathogenic variations **2,624** **Non-pathogenic variations** **196,705**

Multiple sequence alignment

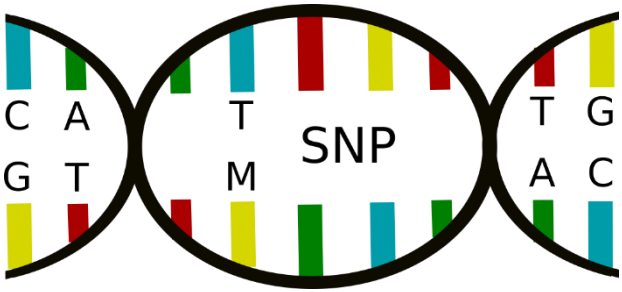
Each line : a protein

Each column: the equivalent position in each protein



Predictor tool

- Compute conservation parameters
- Train a machine-learning model able to classify mutations as pathogenic or not
- Build a web applicaction
  http://lmc.uab.es/tmsnp/

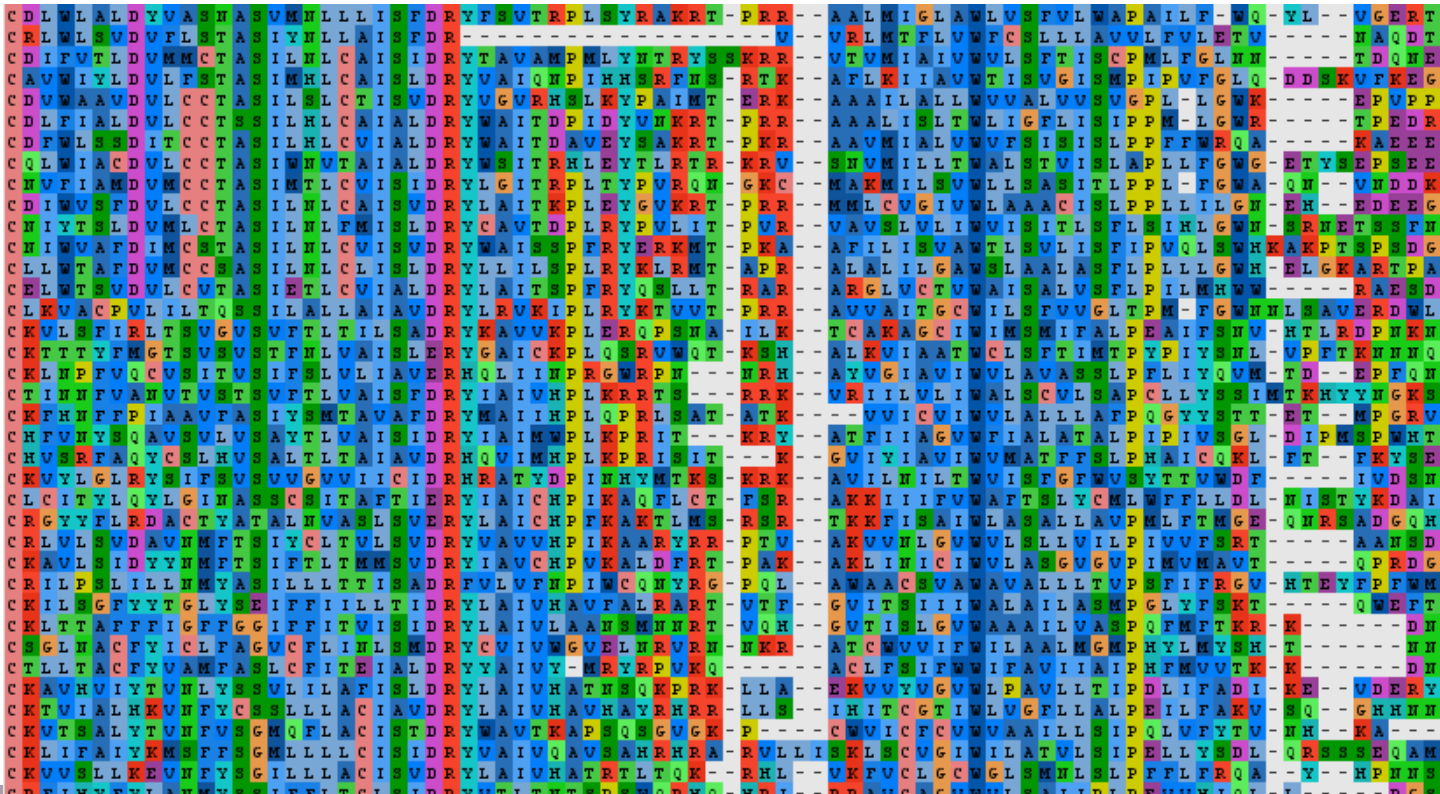


Frequencies of the reference and mutated residue

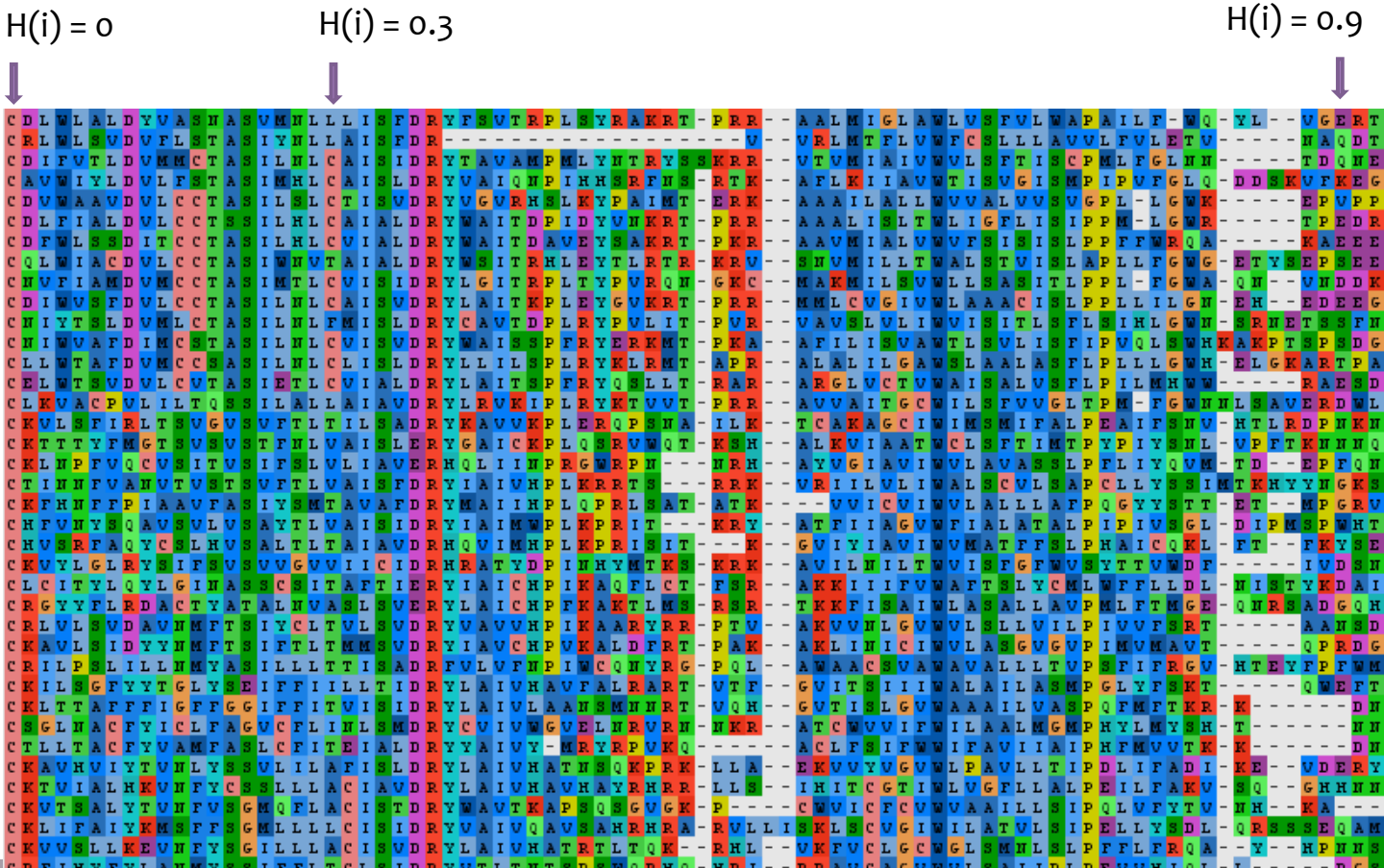
Ex. C (reference) → T (mutated)

$f_{ref} = f_C(i) = 1$
$f_{mut} = f_T(i) = 0$

$f_{ref} = f_C(i) = 0.4$
$f_{mut} = f_T(i) = 0.3$



Entropy (sequence variability, information content)



Substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5																			
R	-6	9																		
N	-2	-3	11																	
D	-5	-7	2	12																
C	1	-8	-2	-7	7															
Q	-3	-2	2	0	-5	9														
E	-5	-6	0	6	-7	1	12													
G	1	-5	-1	-2	-2	-2	-3	9												
H	-3	-4	4	-1	-7	2	-1	-4	11											
I	0	-6	-3	-5	-3	-3	-5	-2	-5	5										
L	-1	-6	-3	-5	-2	-3	-5	-2	-4	2	4									
K	-7	-1	-2	-5	-10	-1	-4	-5	-5	-7	-7	5								
M	-1	-6	-2	-5	-2	-1	-5	-1	-4	3	2	-6	6							
F	-1	-7	-1	-5	0	-2	-5	-2	-2	0	1	-7	0	6						
P	-3	-7	-4	-5	-8	-3	-5	-3	-6	-4	-5	-4	-5	-5	13					
S	2	-6	1	-4	1	-1	-3	1	-2	-2	-2	-5	-2	-2	-3	6				
T	0	-6	-1	-5	-1	-3	-5	-1	-4	-1	-1	-6	0	-2	-4	1	3			
W	-4	-7	-5	-7	-4	1	-7	-5	-3	-4	-3	-8	-4	0	-6	-5	-7	11		
Y	-3	-6	2	-4	-1	0	-2	-3	3	-3	-2	-4	-2	4	-5	-2	-3	1	11	
V	1	-7	-3	-5	-2	-3	-5	-2	-5	3	1	-8	1	-1	-4	-2	0	-4	-3	4

Log odds of finding two amino acids aligned in a sequence alignment

PHAT 75/73 matrix

Statistical methods

Dataset

- 4V (frequencies, entropy, similarity)
- 6V (+ amino acid types)
- 8V (+ proteins and families)

Supervised learning

- Random forest
- Support vector machines
- Gradient Boosting (XGBoost)

- Internal validation (5-fold cross-validation)
- External validation (20% test set)

Random Forest 8V was selected as the final model

Predictor types	Method	Sensitivity	Specificity	MCC	Coverage	Accuracy
Specific for membrane proteins	**TMSNP** (0.95 confidence)	0.90	0.86	0.76	0.38	0.88
	TMSNP (0.90 confidence)	0.86	0.82	0.68	0.58	0.84
	TMSNP (0.80 confidence)	0.81	0.75	0.56	0.86	0.78
	Pre-MutHTP (0.95 confidence)	0.96	0.54	0.56	0.76	0.64
	Pre-MutHTP (0.90 confidence)	0.96	0.53	0.55	0.76	0.67
	Pre-MutHTP (0.80 confidence)	0.96	0.53	0.56	0.76	0.71
Non-specific for membrane proteins	**Polyhen-2**	0.93	0.35	0.35	1	0.64
	SIFT	0.88	0.52	0.42	1	0.70

Applicability domain

Sensitivity: ability to correctly detect pathogenic mutations (TP rate)
Specificity: ability to correctly detect non-pathogenic mutations (TN rate)
MCC = Matthews correlation coefficient; combines sensitivity and specificity

Published online 23 February 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 **1**
doi: 10.1093/nargab/lqab008

TMSNP: a web server to predict pathogenesis of missense mutations in the transmembrane region of membrane proteins

Adrián Garcia-Recio[1,2,†], José Carlos Gómez-Tamayo[3,†], Iker Reina[1], Mercedes Campillo[1], Arnau Cordomí[1,4,*] and Mireia Olivella[2,4,*]

[1]Laboratori de Medicina Computacional, Facultat de Medicina, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain, [2]Bioinformatics and Medical Statistics Group, Facultat de Ciències i Tecnologia, UVIC-UCC, 08500 Vic, Barcelona, Spain, [3]Pharmacoinformatics Group, Research Program on Biomedical Informatics (IMIM/UPF), 08003 Barcelona, Spain and [4]Bioinformatics Department, ESCI-UPF, 08003 Barcelona, Spain

Received May 18, 2020; Revised December 16, 2020; Editorial Decision January 25, 2021; Accepted January 27, 2021