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 St ructure%2BKey%2BLabelled.pn NoBB.png

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

Major groove

DNA to RNA to Protein



CC**GGTAGA**CGAG GGUAGA GR GATCGACTAGTCG ATAGTAGCTACG 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





Tertiary structure

Side chain

F



AGPLRTAWIVALACLDQNVLA

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/



Actin and

myosin

Stator

DNA sequence

Protein sequence

Protein structure

Protein function



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





Will mutations have a functional effect?



GCTAGCTATCTA**TAT**GGGATATGA

UAU: Y



¹The one letter symbol of amino acids.

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

Cost per Genome





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?







Comparing sequences is very cheap!

KNOWN QMKEDAKGKSEEELAECFRIFDRNMDGYIDAEELAEI

UNKNOWN QMKEDAKGKSEEELAECFRIFDRNMDG**F**IDAEELAEI

Compare related proteins

CMK**T**DSKGKSEDELSDLFRMFDKNADG**Y**IDAEELADL QMK**E**DAKGKSEEELAECFRIFDRNMDG**Y**IDAEELAEI QMK**A**DAKGKSEEELAECFRMFDKNADG**Y**IDLDELADV QMK**S**DAKGKSEEELAECFRIFDRNMDG**Y**IDAEELAEI

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

A protein family



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 aggregation database

Genome database (140,000 individuals



CTGATGGTATGGGGCCAAGAGATA AGGTACGGCTGTCATTAGAC AGGGCTGGGATAAAAGTCAGGGC/ CATGGTGCATCTGACTCCTGAGGA CAGGTTGGTATCAAGGTTACAAGA(GCACTGACTCTCTCTGCCTATTGG 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

CDLWLALDYVASHAS	SVMNLLLISFDF	YFSVTRPLSYRAKRT-PRR-	- A A L M I G L A W L V S	FVLWAPAILF-WQ) - YL VGER 1
CRLWLSVDVFLSTAS	S I Y N L L A I S F D P	<mark>v</mark> -	- <mark>V R</mark> L M T F L V W F C S	LLLAVVLFVLETV	7 N A Q D 1
CDIFVTLDVMMCTA:	S I L N L C A I S I D P	YTAVAMPMLYNTRYSSKRR -	- VTVMIAIVWVLS	FTISCPMLFGLNP	г т р q н в
CAVWIYLDVLFSTAS	S I M H L C A I S L D P	YVAIQUPIHHSRFUS-RTK-	AFLKIIAVWTIS	VGISMPIPVFGLQ	- D D <mark>S K V F K E</mark> (
CDVWAAVDVLCCTAS	5 I L S L C T I S <mark>V D P</mark>	YVGVRHSLKYPAIMT-ERK-	AAAILALLWVV	L V V S V G P L - L G W F	(E P V P F
CDLFIALDVLCCTS	ILHLCAIALDP	YWAITDPIDYVNKRT-PRR-	- A A A L I <mark>S</mark> L T W L I G	FLISIPPM-LGWP	T P E D F
CD FWL S S D I T C C T A S	ILHLCVIALDP	YWAITDAVEYSAKRT-PKR-	- A A V M I A L V W V F S	ISISLPPFFWRQA	<mark>K A</mark> E E B
CQLWIACDVLCCTAS	SIWNVTAIALDP	YWSITRHLEYTLRTR - KRV -	- SHVMILLTWALS	TVISLAPLLFGWG	- E T <mark>Y S</mark> E P S E B
CHVFIAMDVMCCTAS	5 I M T L C V I S I D P	YLGITRPLTYPVRQM-GKC-	- MAKMILSVWLLS	ASITLPPL-FGWA	- o n v n d d n
CD I W V S F D V L C C T A S	S I L N L C A I S V D P	YLAITKPLEYGVKRT-PRR-	- MMLCVGIVWLA	ACISLPPLLILGP	I - E H E D E E (
CNIYTSLDVMLCTAS	S I L N L F M I S L D F	YCAVTDPLRYPVLIT-PVR-	VAVSLVLIWVIS	ITLSFLSIHLGWP	I - SRNETSSFI
CHIWVAFDIMCSTAS	5 I L N L C V I S V D P	YWAISSPFRYERKMT-PKA-	AFILISVAWTLS	VLISFIPVQLSWH	K A K P T S P S D 0
CLLWT AFDVMCCSAS	S I L N L C L I S L D P	YLLILSPLRYKLRMT - APR -	- ALALILGAWSLA	ALASFLPLLLGWH	I - E L G K A R T P J
CELWTSVDVLCVTAS	S I E T L C V I A L D P	YLAITSPFRYQSLLT-RAR-	ARGLVCTVWAIS	ALVSFLPILMHWW	RAESI
CLKVACPVLILTQS	I L A L L A I A V D P	YLRVKIPLRYKTVVT - PRR -	AVVAITGCWILS	FVVGLTPM-FGWP	INLSAVERDWI
CKVLSFIRLTSVGVS	SVFTLT IL SADE	YKAVVKPLERQPSHA-ILK-	TCAKAGCIWIMS	MIFALPEAIFSNV	/ - HTLRDPNKI
CKTTTYFMGTSVSVS	S T F N L V A I S L E F	YGAICKPLQSRVWQT-KSH-	- ALKVIAATWCLS	FT IMTPYPIYSNL	<mark>V P F T K</mark> HNN (
CKLHPFVQCVSITVS	SIFSLVLIAVEP	HQLIINPRGWRPNNRH-	- A Y V G I A V I W V L A	VASSLPFLIYQVM	I - T D E P F Q I
CTINNFVANVTVST:	SVFTLVAISFDF	YIAIVHPLKRRTSRRK-	V R I I L V L I W A L S	CVLSAPCLLYSSI	M T K H Y Y N G K S
CKFHNFFPIAAVFAS	S I Y S M T A V A F D F	YMAIIHPLQPRLSAT-ATK-	<mark>VVICVIWVL</mark>	LLLAFPQGYYSTT	- E T M P G R 1
CHFVHYSQAVSVLVS	S A Y T L V A I S I D P	YIAIMWPLKPRIT KRY-	- A T F I I A G V W F I A	LATALPIPI <mark>V</mark> SGL	<mark>D I P M S P W H</mark> T
CHVSRFAQYCSLHVS	3 A L T L T A I A V D P	HQVIMHPLKPRISIT K -	- GVIVIAVIWVM A	TFFSLPHAICQKL	FT FKYSI
CKVYLGLRYSIFSVS	S V V G V V I I C I D P	HRATYDPINHYMT <mark>KS</mark> -KRK-	- A V I L <mark>H</mark> I L T W V I <mark>s</mark>	F G F W V S Y T T V W D F	IVD 51
CLCITYLQYLGINAS	S S C S I T A F T I E F	YIAICHPIKAQFLCT - FSR -	- A <mark>K K</mark> I I I F V W A F I	<mark>SLYCMLWFFLLD</mark> L	NISTY <mark>KDA</mark> I
CRGYYFLRDACTYAT	Г А L И V А S L S V E F	YLAICHPFKAKTLMS-RSR-	- T <mark>K K</mark> F I <mark>S</mark> A I W L A S	ALLAV <mark>PMLFTMGE</mark>	: - Q N <mark>R S A D G Q </mark> B
CRLVLSVDAVNMFTS	3 I Y C L T <mark>V</mark> L S <mark>V D F</mark>	YVAVVHPIKAARYRR - PTV -	- A <mark>K V V N</mark> L <mark>G V W V</mark> L <mark>S</mark>	LLVILPIVVFSRT	<mark>aans</mark> i
CKAVLSIDYYNMFTS	3 I F T L T M M S <mark>V D F</mark>	YIAVCHPVKALDFRT-PAK-	- A <mark>K</mark> L I <mark>N I C</mark> I W V L A	S G V G V P I M V M A V T	QPRD(
CRILPSLILLMMYAS	SILLLTTI <mark>SA</mark> DF	FVLVFNPIWCQNYRG-PQL-	- <u>A W A A <mark>C S</mark> V A W A V A</u>	LLLT <mark>VPSFIFRG</mark> V	7 - H T E Y F P F W M
CKILSGFYYTGLYSI	E I F F I I L L T I D F	YLAIVHAVFAL <mark>RAR</mark> T-VTF-	- GVITSIIIWAL 🛛	ILASMPGLYFS <mark>K</mark> T	Q W E F 1
CKLTTAFFFIGFFGG	5 I F F I T <mark>V</mark> I S I D F	YLAIVLААП SMНИ <mark>R</mark> T - <mark>VQH</mark> -	- <mark>G V T I S L G V</mark> W A A A	ILVASPQFMFTKR	L - K D I
CSGLHACFYICLFAG	GVCFLINLSMDF	YCVIVWGVELNRVRN-NKR-	- A T C W V V I F W I L P	ALMGMPHYLMY <mark>S</mark> H	ι – <u>τ</u> – – – – – – <u>π</u> 1
CTLLTACFYVAMFAS	S L C F I T E I A L D F	<u> </u>	- A C L F <mark>S</mark> I F W W I F A	VIIAIPHFMVVT K	(- <mark>K D 1</mark>
CKAVHVIYTVHLYSS	S V L I L A F I S L D F	YLAIVHATNSQ <mark>KPRK</mark> -LLA-	- EKVVYVGVWLP	VLLTIPDLIFAD I	[- KE VDER 3
CKTVIALHKVHFYC	S S L L L A C I A V D F	YLAIVHAVHAY <mark>R</mark> H <mark>RR</mark> - LL <mark>S</mark> -	IHITCGTIWLVG	FLLALPEILFAKV	7 - <mark>5 q</mark> <mark>6 н н п 1</mark>
CKVTSALYTVNFVSG	GMQFLACISTDF	YWAVTKAPSQSGVGK-P	- CWVICFCVWVA2	ILLSIPQLVFYTV	7 - N H K A
CKLIFAIYKMSFFS	GMLLLL <mark>CISID</mark> F	YVAIVQAVSAHRHRA-RVLL	I <mark>S K</mark> L <mark>S C V G I W</mark> I L 🛛	TVLSIPELLY <mark>SD</mark> L	Q <mark>R</mark> S S S E Q A M
CKVVSLLKEVNFYS	FILLLACISVDE	YLAIVHATRTLTOKRHL-	- VKFVCLGCWGL	MHLSLPFFLFR02	<mark>ү</mark> <mark>нр</mark> ии :

Pfam

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





Frequencies of the reference and mutated residue

Ex. C (reference) \rightarrow T (mutated)

$$f_{ref} = f_C(i) = 1$$

 $f_{mut} = f_T(i) = 0$
 $f_{mut} = f_T(i) = 0.3$

						*																						
сп	JL WL	. A L D	Y V A	S II A	SVM	N L L L	. I <mark>S</mark> I	FDR	Y F S	VT	R P L	S Y R	A K	R T -	PRI	R	A A L M	IGL	N W L V	STV	L W A	P A I	LF-	W Q	- YL	🗸	GE	R T
C F	🧧 L 💟 L	. <mark>s v d</mark>	VFL	ST A	SIY	N L L A	ISI	FDR								7 1	V <mark>R</mark> L M	TFL	7 W F C	SLL	LAV	VLF	VLE	ТV		1	λQ	DT
СІ	JIFV	TLD) V M M	CTA	SIL	H L C A	IS	IDR	ҮТ А	VAN	IP MI	LYN	TR	YSS	KRI	R	νтνм	IAI	7 W V L	SFT	ISC	PML	FGL	n n		T	DQI	IT E
C 2	A V W I	(YL <mark>D</mark>	VLF	ST A	SIM	HLCA	ISI	LDR	ΥVA	IQI	I P I I	H H S	RF	17 <mark>S</mark> -	RTI	K	A F L K	IIA	7 W T I	SVG	ISM	PIP	VFG	ΓQ	- D D	S K V	FKI	E G
СІ	1 V W A	1 A V D	J V L C	СТА	SIL	S L C T	IS	V D R	YVG	VR	I S L	K Y P	λI	мт-	ERI	K	AAAI	LALI	. W V V	ALV	VSV	GPL	- L G	WK		E	РVІ	ΡΡ
СІ	JLFI	i a l D	JVLC	СТЗ	SIL	HLCA	IAI	LDR	YWA	ITI	PI	DYV	N K	R T -	PRI	R	AAAL	ISLI	WLI	GFL	ISI	ррм	- L G	WR		<u>T</u>	ΡEI	D R
СІ	J F W L	. <mark>S S D</mark>	JITC	СТА	SIL	HLCV	7 I A I	LDR	YWA	ITI	A V I	EYS	A K	r T -	PKI	R	AAVM	IAL	/WVF	SIS	ISL	РРГ	FWR	Q A		K	A E I	ΕE
C Q	I L M I	I A C D	VLC	СТА	SIW	И V Т А	IAI	LDR	YWS	ITI	R H L I	EYT	LR	T R -	KR	/	з п 🗸 м	ILL	WAL	STV	ISL	APL	LFG	WG	- E T	YSE	PSI	ΕE
c I	IVFI	AMD	UM C	CTA	SIM	TLCU	7 I S 3	IDR	Y L G	ITI	PL	ΓΥΡ	VR	Q 11 -	GK	c	MAKM	ILS	WLL	SAS	ITL	PPL	- F G	WA	- Q 11	0	ום ת	D K
C I		SFD	ULC	СТА	SIL	N L C A	IS	D D R	YL A	ITI	PLI	EYG	VK	RT-	PRI	R	MMLC	VGI	WLA	AAC	ISL	PPL	LIL	G N	- E H	E	DEI	E G
c r	IIYT	SLD	JVML	CTA	SIL	N L F M		LDR	Y C A	UTI	PL	R Y P	VL	I T -	PVI	R	VAVS	LVL		SIT	LSF	LSI	HLG	w n	- 5 R	NET	SS	гn
C I	TIWU	AFD	IMC	STA	SIL	игсо	1 1 2 1	D R	YWA	ISS	P F	RYE	RK	мт-	PK	A	AFIL	ISV	WTL	SVL	ISF	IPV	QLS	W H	KAK	PTS	PSI	DG
CI	. L W T	AFD	UM C	CSA	SIL	пгсг	. 1 5 1	LDR	YLL	IL	PL	RYK	LR	M T -	A P I	R	ALAL	ILG	WSL	AAL	ASF	LPL	LLG	WН	- EL	GKA	RT	PA
CE	LWT	SVD	UVL C	VTA	SIE	TLCV		LDR	YLA	IT	P P	RYQ	SL.	L T -	RAI	R	ARGL	VCT	/ W A 1	SAL	V S F	LPI	LMH			R	AE	3 0
C I		СРО		TQB	SIL	ALLA		D R	YLR	VK.		RYK	TV	V T -	PRI	R		ITGO	IL	STV	GL	TPM	- r G	W M	HL S	AVE	RD	L
C P		I I I	LTS		SVP	TLTI	LSI	A D R	YKA		PLI		P S	1 A -			TCAR	AGC		SM 1	r a L	PEA	1 1 5		- H T	LKD	РИ	
0 1									T G R	1				U I -				100				PIP	TTO	11 L		I I K		a u
								P D D	n u L				р <u>и</u>					TUT				PIL	TTU			U U U	PI	U 10
					- T										X K I							PCL		D 1				
					0 1 1 0 1 V				U T A			U P E		<u></u> .			A T F I	TAC				PUG	TUP	C 1		D M S	P G	K V
				TWO		T T T 1			H O U					T T -				1 1 0 0					TCO	7 7				
		GIT		FRU	3 1 1				HDA	TV		THV	L D	V 9 -	V D T			TTTT		S F G		9 V T	TUN	D F			V D	
c 1	CTT	VL C	VI.G	THA	5 5 C	SITA	T T	TEP	V T A	TC	T D T		F L	ст-	F S I		AKKI	TTES	7 W A F	TSL	V C M	LWF	FLL	D L	- 17 T	S T V	K D	A T
c I	GVY	T L F	D A C	TYA	TAL	H V A S	T. S	TER	Y L A	TC	TPF	K A K	TL	M S -	B S I		TKKF	TRA	T W T. A	SAL		PMT.	F T M	GE	- 0 1	B S A	n G	ОН
C F		SVI	AVE	MFT	SIY	CLTV	LS	DR	YVA	VV	IP I	K A A	RY	RR-	РТ	7 1	AKVV	HLG	V W V L	SLL	VIL	PIV	VFS	RT		3	AN	S D
C F	X A V I	SIT	YYP	MTT	SIF	TLTM	IMS	DR	YIA	vc	I P V	KAL	DF	R T -	PA	K	AKLI	піс	WVL	ASG	VGV	PIM	VMA	VΤ		0	PR	DG
C F	RILP	SLI	LLN	AYM	SIL	LLTT	IS	ADR	FVL	VFI	I P I	w c o	πч	RG-	POI		AWAA	CSV	WAV	ALL	LTV	PSF	IFR	GV	- нт	EYF	PF	wм
C I	KIL!	GFY	YTG	LYS	EIF	FIIL	LT	IDR	YLA	IVI	I A V	FAL	RA	R T -	VTI	F	GVIT	SIII	WAL	AIL	ASM	PGL	YFS	КТ		0	WE	г т
СР	LTT	AFF	FFIG	FFG	GIF	FITV	7 I S 3	IDR	Y L A	IVI	AA	и з м	เ ท ท	R T -	vqı	H	GVTI	SLG	7 W A A	AIL	VAS	PQF	МГТ	KR	- <mark>K</mark> -			рπ
c s	S G L D	ACI	r y I C	LFA	GVC	FLIN	LS	M D R	YCV	IVI	GV	ELN	RV	<mark>в 11</mark> –	ики	R	АТС М	VVII	WIL	AAL	MGM	рнү	LMY	S H	-т-			пп
ст	C L L T	ACI	YVA	МГА	SLC	FITE	IAI	LDR	ΥYA	IV	Z - M	R Y R	ΡV	K Q -			ACLF	SIF	WIF	λVΙ	ΙΑΙ	р н г	MVV	тк	- <mark>K</mark> -			D N
C P	K A V F	ΙνΙγ	TVN	LYS	SVL	ILAF	ISI	LDR	YL A	IVI	I A T I	n s q	K P	R K -	L L .	A	ε <mark>κ</mark> νν	YVGV	/WLP	AVL	LTI	PDL	IFA	DI	- <mark>K E</mark>	🗸	DE	R Y
C I	ст V ј	ALH	к <mark>к</mark> v и	ГҮС	SSL	LLAC	IA	V D R	YL A	IVI	τ A V I	нач	R H	R R -	LL	5	IHIT	CGT	I W L V	GFL	LAL	P E I	LFA	ĸν	- 5 Q	G	нн	n n
C I	CVT 5	ALY	ат у п	FV S	GMQ	FLAC	IST	T D R	Y W A	VT	K A P	s Q s	GV	G K -	P		CWVI	CFC	V W V A	AIL	LSI	PQL	VFY	ΤV	- <mark>п</mark> н	K	Α - ·	
C P	CLII	AIY	KM S	Г Г З	GML	LLLC	IS	IDR	Y V A	IV	A V	S A H	r R	R A -	r v i	LLI	S <mark>K</mark> L S	CVG	IWIL	ATV	LSI	PEL	LYS	DL	- Q R	SSS	ΕQ	A M
C I	K V V S	LLF	E V 1	FYS	GIL	LLAC	IS	V D R	YL A	IVI	I A T	R T L	ΤQ	K	R H I	L	VKFV	CLGO	WGL	з м п	L <mark>S</mark> L	РГГ	LFR	Q A	Y	H	P N I	IT S
C . T	TT T T	1 2 2 2	7 T 3 17	1 M 32 Q	O T T	TT T	T 0	TDD	N2 82 T	T T 1	T T 9 1			¥ 0 -	W D	T	D D A V	C 2 C 3	2 TA 5 2 T	C A T	TDT	D E 12	52 H T	0.7				0 0

Entropy (sequence variability, information content)

H(i) = 0	H(i) = 0.	3			H(i) = 0.9
l	Ļ				Ļ
C D L W L A L D Y V A S C R L W L S V D V F L S	N A S <mark>V M</mark> N L L L I S F D I T A S I Y N L L A I S F D I	YFSVTRPLSYRAKRT-	<mark>P R R</mark> A A L M I G L A W L <mark>V</mark> V <mark>R</mark> L M T F L V W F	VSFVLWAPAILF-WQ CSLLLAVVLFVLETV	- YL VGERT NAQDT
C D I FVTLDVMMC C AVWIYLDVLF C D VWAAVDVLCC	TASILNLCAISIDI TASIMHLCAISLDI TASILSLCTISUDI	YTAVAMPMLYNTRYSS YVAIQNPIHHSRFNS- YVGVPHSLKYPAIMT-	K R R O T O M I A I O W U R T K A F L K I I A V W T E R K A A A I L A L L W U	LSFTISCPMLFGLHH ISVGISMPIPVFGLQ VALVVSVGPL-LGWK	T D Q N E - D D <mark>B K V F K E</mark> G
C D L F I A L D V L C C C D F V L S S D I T C C	T S S I L H L C A I A L D F T A S I L H L C V I A L D F	YWAITDPIDYVNKRT- YWAITDAVEYSAKRT-	P R R A A A L I <mark>S</mark> L T W L P K R A A V M I A L V W V	IGFLISIPPM-LGWR FSISISLPPFFWRQA	<mark>T P E D R</mark> <mark>K A E</mark> E E
C Q L W I A C D V L C C C N V F I A M D V M C C C D I W V M F D V I C C	T A S I W H V T A I A L D H T A S I M T L C V I S I D H	YWS ITRHLEYTLRTR YLG ITRPLTYPVRQN-	KRVSHVMILLTWA GKCMAKMILSVWL BBBMMICVCIVI	L S T V I S L A P L L F G W G L S A S I T L P P L - F G W A	- E T Y S E P S E E - Q II VII D D K - E N E D E E C
CHIYTSLDVMLC CHIWVAFDIMCS	TASILNLFMISLDF TASILNLCVISVDF	Y C A V T D P L R Y P V L I T - Y W A I S S P F R Y E R K M T -	PVR VAVSLVLIWU PKA AFILISVAWT	ISITLSFLSIHLGWN LSVLISFIPVQLSWH	- SRNETSSFN KAKPTSPSDG
C L L W T A F D V M C C C E L W T S V D V L C V C L V A C P V L L T	S A S I L N L C L I S L D I T A S I E T L C V I A L D I	YLLILSPLRYKLRMT- YLAITSPFRYQSLLT-	APRALALILGAWS RARARGLVCTVWA	L A A L A S F L P L L L G W H I S A L V S F L P I L M H W W	- ELGKARTPA RAESD
C K V L S F I R L T S V C K T T T Y F M G T S V	GVSVFTLTILSAD SVSTFNLVAISLE	Y K A V V K P L E R Q P S N A - Y G A I C K P L Q S R V W Q T -	I L K T C A K A G C I W I K S H A L K V I A A T W C	MSMIFALPEAIFSIV LSFTIMTPYPIYSNL	- HTLRDPNKN - VPFTKNNNQ
C K L H P F V Q C V S I C T I H H F V A H V T V C K F H H F F F L A A V T	ITVSIFSLVLIAVE STSVFTLVAISFD FDSVFTLVAISFD	HQLIINPRGWRPN YIAIVHPLKRRTS	NRH AYVGIAVIWU RRK VRIILVLIWA	LAVASSLPFLIYQVM LSCVLSAPCLLYSSI	- T D E P F Q N M T K H Y Y N G K S
CHFVHYSQAVSV CHVSRFAQYCSL	LVSAVTLVAISIDE HVSALTLTAIAVDE	YIAIMWPLKPRIT HQVIMHPLKPRISIT-	KRYATFIIAGUWF KGVIYIAVIWU	I ALLATALPIPIUS GL MATFFSLPHAICQK	- D I P M S P W H T - F T F K Y S E
C K V Y L G L R Y S I F C L C I T Y L Q Y L G I	SVSVVGVVIICIDI INASSCSITAFTIE	H R A T Y D P I N H Y M T K S - Y I A I C H P I K A Q F L C T -	KRK AVILNILTWU FSR AKKIIIFUWA	ISFGFWVSYTTVWDF FTSLYCMLWFFLLDL	IVDSN - NISTYKDAI
CRUVLSVDAVIM CRLVLSIDYYIM	IFTSIVCLTVLSVDI IFTSIFTLTMMSVDI	YVAVVHPIKARTLUB- YVAVVHPIKAARYRR- YIAVCHPVKALDFRT-	PTVAKVVHLGVWU PAKAKLIHICIWU	L S L L V I L P I V V F S R T L S G V G V P I M V M A V T	
CRILPSLILLNM CKILSGFYYTGL	IYASILLLTTISADI YSEIFFIILLTIDI	FVLVFNPIWCQNYRG- YLAIVHAVFALRART-	PQLAWAACSVAWA VTFGVITSIIIWA	VALLLTVPSFIFRGV LAILASMPGLYFSKT	- <u>H T E Y F P F W M</u> Q W E F T
C K L T T A F F F I G F C S G L H A C F Y I C L C T L L T A C F Y V A M	FAGUCFLINLSMD FAGUCFLINLSMD FASLCFITEIALDE	Y L A I V L A A II S M II II R T - . Y C V I V W G V E L II R V R II - . Y Y A I V Y - M R Y R P V K Q -	NKR ATCWVVIFWI ACLFSIFWWI	LAALMGMPHYLMYSH FAVIIAIPHFMVVT	- к ли - т ли - к ли
C K A V H V I Y T V H L C K T V I A L H <mark>K</mark> V H F	Y S S V L I L A F I S L D F Y C S S L L L A C I A V D F	YLAIVHATHSQKPRK- YLAIVHAVHAYRHRR-	L L A E K V V Y V G V W L L L <mark>S</mark> I H I T C G T I W L	PAVLLTIPDLIFADI VGFLLALPEILFAK	- ке у реку - в q сннии
C K V T S AL Y T V N F C K L I F A I Y K M S F C K V V S L L K E V N F	FSGMLLLLCISIDE YSGILLLACISVDE	Y W A V T K A P S Q S G V G K - . Y V A I V Q A V S A H R H R A - Y L A I V H A T R T L T Q K	P C W V I C F C V W R V L L I S K L S C V G I W I R H L V K F V C L G C W G	LATULSIPQLVFYTU LATULSIPELLYSDL LSMNLSLPFFLFRQA	- U H K A - Q R S S S E Q A M Y H P N N S

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

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

PHAT 75/73 matrix

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 Reina¹, Mercedes Campillo¹, Arnau Cordomí^{(1,4,*} and Mireia Olivella^{(1,2,4,*})

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

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