

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Pier Luigi Martelli

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Precentral Material 1 (PCM1): predicting the effects of missense changes on a zebrafish model

Giulia Babbi,
Pier Luigi Martelli,
Castrense Savojardo,
Rita Casadio



Biocomputing Group
University of Bologna
www.biocomp.unibo.it



PCM1 challenge

- **Input:** 38 missense variants within the Pericentral Material 1 (PCM1) gene. ***Q15154 (PCM1_HUMAN)***
- Variants were assessed in a zebrafish model for the impact on brain development
- **Challenge:** provide predictions about the impact of each variant on brain development
 - Three classes: benign, hypomorphic and pathogenic
- **Our approach:** Disease Index Matrix (Casadio *et al.*, 2011)

• *Pd index: Learning from data bases*



Single Amino acid Polymorphism (SAP) data sets

SAP-Data Base	SAPs *			Proteins *
	<i>Disease related</i>	°Neutral	Total	
dbSNP	-	7,417	7,417	4,781
UniProt	10,322 (OMIM)	-	10,322	837
SAP-UniProt+dbSNP (150 SAP types)	10,322	7,417	17,739	5,380

Disease related SAPs (Single Aminoacid Polymorphisms) of globular proteins are derived from UniProtKB (release 2010_04/23 March 2010) considering only mutations that are OMIM annotated and excluding those related to somatic cancer. The set of neutral mutations comprises all the missense mutations that were retained from dbSNP (11/09/2010 RELEASE, Build 132, <http://www.ncbi.nlm.nih.gov/projects/SNP/>), provided that the minor allele frequency is >5% in each and all (12) populations. The SAP-Final dataset is obtained by merging both UniProt and dbSNP mutations (SAP-UniProt+dbSNP)

*Figures are given as numbers of SAPs and proteins, respectively.

Computing a disease index: **Pd**



SAP type ($X \rightarrow Y$) (X and Y are the wild type residue and the mutated residue, respectively) the associated disease probability (or *disease index* $Pd(X \rightarrow Y)$) is :

$$Pd(X \rightarrow Y) = Nd(X \rightarrow Y) / N(X \rightarrow Y)$$

$Nd(X \rightarrow Y)$ is the number of $X \rightarrow Y$ mutations conducive to a any kind of disease

$N(X \rightarrow Y)$ is the total number of mutations $X \rightarrow Y$

	G	A	V	P	L	I	M	F	W	Y	S	T	C	N	Q	H	D	E	K	R
G		.64 91(73)	.82 239(134)						.78 33(30)		.66 334(221)		.81 108(59)				.82 286(144)	.77 233(142)		.76 64(37)
A	.47 58(54)		.45 383(325)	.83 39(113)							.33 119(113)	.43 415(344)					.57 181(91)	.51 65(58)		
V	.62 75(69)	.33 178(162)			.42 117(105)	.16 248(239)	.50 223(185)	.69 54(47)									.83 44(38)	.83 31(25)		
P		.43 68(67)			.47 443(358)						.48 224(184)	.55 84(74)			.40* 38(38)	.51* 47(42)				.73 92(75)
L			.43 114(101)	.77 473(295)		.14 46(45)	.21 66(62)	.50 157(157)	.57 81(30)		.56 71(68)			.77 48(37)	.63 25(22)					.86 134(106)
I			.16 191(185)		.33 38(37)		.47 84(89)	.87 74(28)			.72 43(37)	.62 262(154)	.90 56(49)						.78 6(5)	.65 13(12)
M			.47 146(123)		.34 43(42)	.45 87(79)					.41 143(128)								.73 29(28)	.78 41(39)
F				.78 33(29)	.59 137(111)	.62 25(24)				.31 12(12)	.81 50(83)		.76 47(41)							
W	.74* 38(33)				.75 81(15)						.88 23(21)		.92* 61(46)							.66 138(118)
Y							.41 10(10)				.77 31(26)	.80 246(166)	.75 28(26)	.62 91(84)	.81 45(37)					
S	.34 78(74)	.19 58(90)	.55 168(147)	.56 138(126)	.39 62(57)		.45 149(154)	.55 7(7)	.51 43(41)		.24 78(99)	.47 68(81)	.28 163(158)							.62 145(90)
T		.27 289(199)	.58 75(72)		.50 215(182)	.49 178(168)					.22 49(67)		.42 55(51)						.52 38(34)	.67 45(41)
C	.78 41(39)						.85* 82(59)	.82* 51(35)	.83 266(125)	.74 78(51)										.75* 208(140)
N						.84 34(29)			.72* 28(28)	.49 218(189)	.56 52(51)			.49 35(35)	.46 81(77)		.59 121(107)			
Q			.75 65(61)	.62 25(25)									.37 117(105)		.44 55(52)	.43 84(87)	.37 186(172)			
H			.68 58(51)	.76 28(25)					.58 113(91)			.54 23(22)	.45 69(64)		.69 35(35)		.50 181(165)			
D	.66 147(125)	.59 40(36)	.79 81(66)						.73 103(79)		.52 281(238)		.69 74(64)		.37 112(110)					
E	.53 133(114)	.51 84(80)	.67 44(39)										.45 86(75)		.35 123(116)		.61 47(33)			
K					.68 18(17)	.70 6(6)					.52 36(35)	.51 114(99)	.43 49(47)			.48 157(142)		.32 118(115)		
R	.68 286(167)	.84 774(135)	.77 127(101)	.47* 15(15)	.21 15(15)		.76 412(248)	.60 112(108)	.71 44(41)	.75 495(328)	.59 318(298)	.64 472(358)						.21 169(142)		

The Disease Index values (Pd) for each SAP type

The darker is the color, the higher the Pd value.

Wild type residues are shown on the left and target residues at the top.

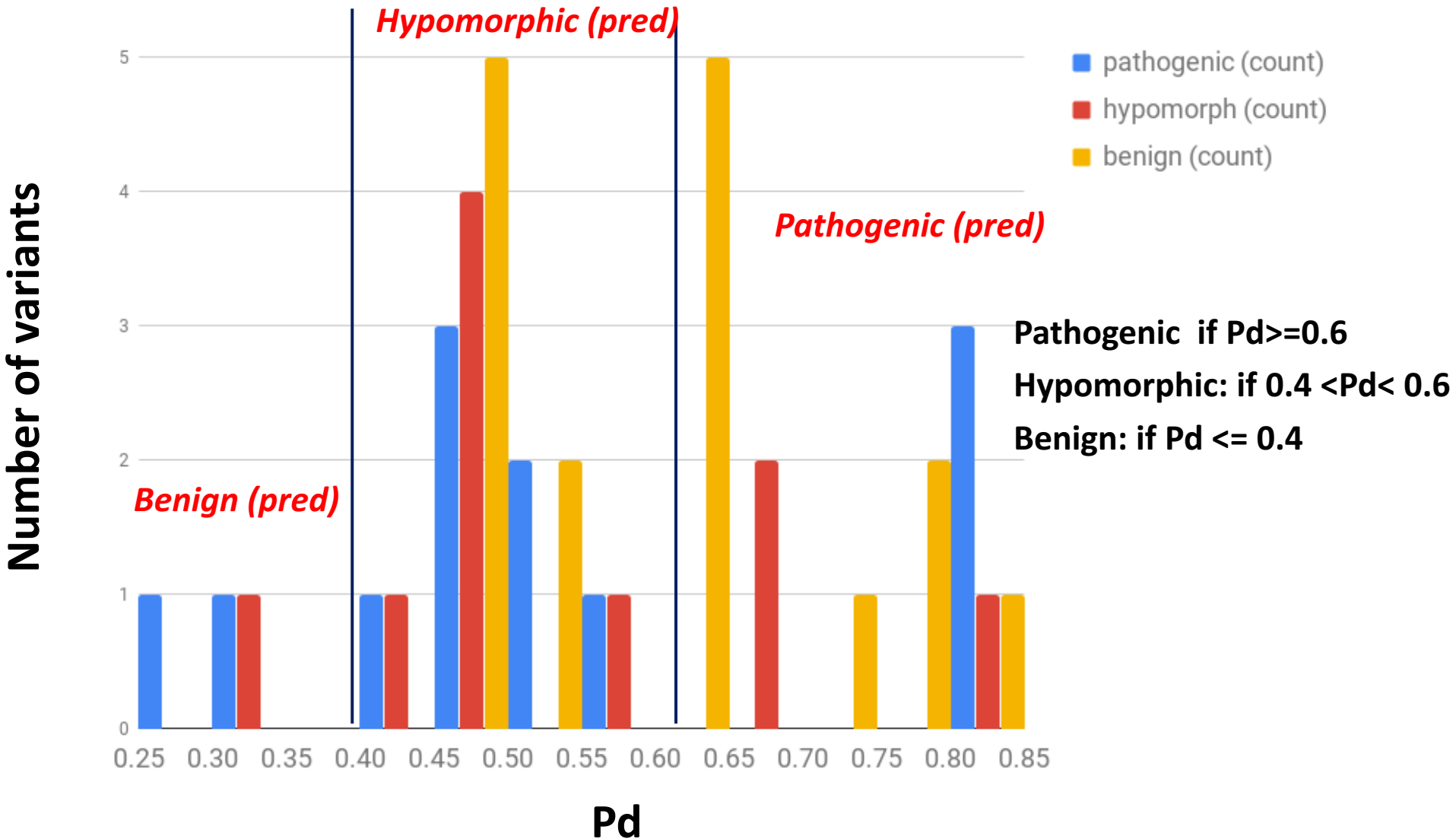
For each mutation type the corresponding number of mutations and, between round brackets, the number of proteins in which mutations are annotated, are shown. Pd is computed for the 150 SAP types due to SNPs. Squares comprising mutations within the same residue group are coloured in red, blue, green and yellow, respectively. All the other off-diagonal squares comprising mutations included in different groups are shown as a combination of the four colours



G	A	V	P	L	I	M	F	W	Y	S	T	C	N	Q	H	D	E	K	R
	.64 94(75)	.82 235(134)						.78 33(39)		.66 334(221)		.81 194(59)				.82 288(166)	.77 232(142)		.76 564(307)
.47 58(54)		.45 383(325)	.83 139(113)							.33 119(113)	.43 415(344)					.57 181(91)	.51 65(55)		
.62 77(69)	.33 178(162)			.42 117(105)	.16 248(239)	.50 223(185)	.69 54(47)									.83 44(38)	.83 31(29)		
	.43 68(67)			.47 443(358)						.48 224(186)	.55 84(74)			.40* 39(38)	.51* 47(42)				.73 92(75)
		.43 114(101)	.77 473(295)		.14 46(45)	.21 66(62)	.50 187(157)	.57 38(19)		.56 71(60)				.77 88(37)	.63 23(22)				.86 134(106)
		.16 191(185)		.33 38(37)		.47 84(89)	.87 34(28)			.72 42(37)	.62 202(154)		.90 56(49)					.78 5(5)	.65 13(12)
		.47 146(123)		.34 43(42)	.45 87(78)					.41 142(129)								.73 29(28)	.78 41(39)
		.78 33(29)		.59 137(111)	.62 25(24)				.31 12(12)	.81 92(83)		.76 47(41)							
.74* 39(33)				.75 16(15)						.88 22(21)		.92* 61(46)							.66 130(118)
							.41 13(13)			.77 31(28)		.80 246(166)	.75 28(26)		.62 91(84)	.81 43(37)			
.34 78(74)	.19 54(50)		.55 168(147)	.56 138(126)	.39 62(57)		.45 169(154)	.55 7(7)	.51 43(41)		.24 70(69)	.47 64(61)	.28 163(156)						.62 145(90)
	.27 289(199)		.58 75(72)		.50 215(182)	.49 179(169)				.22 69(67)			.42 55(51)					.52 38(34)	.67 45(41)
.78 61(39)							.85* 82(59)	.82* 51(35)	.83 266(125)	.74 78(51)									.75* 208(146)
					.84 34(29)				.72* 28(28)	.49 218(189)	.56 32(31)				.49 35(35)	.46 81(77)		.59 121(107)	
			.75 65(61)	.62 25(25)											.37 117(106)		.44 55(52)	.43 99(57)	.37 186(172)
			.68 53(51)	.76 29(25)					.58 112(91)				.54 22(22)	.45 69(66)		.69 35(35)			.50 181(163)
.66 147(125)	.59 48(36)	.79 81(66)							.73 103(79)			.52 281(238)		.69 74(64)		.37 112(110)			
.53 133(114)	.51 64(60)	.67 44(39)											.45 80(75)		.35 123(116)		.61 447(318)		
					.68 18(17)	.70 6(6)					.52 36(35)		.51 114(98)	.43 49(47)		.48 157(142)			.32 118(115)
.68 299(167)			.84 174(135)	.77 127(101)	.47* 15(15)	.21 15(15)		.76 412(299)		.60 122(108)	.71 44(41)	.75 495(328)		.59 518(398)	.64 76(71)			.21 10(10)	

Results on 38 PCM1 variants

Distribution of PD values for pathogenic, hypomorphic and benign variants



Results

Predicted

P H B

Observed

P

4

6

2

H

3

6

1

B

9

7

0

Pathogenic if $Pd \geq 0.6$

Hypomorphic: if $0.4 < Pd < 0.6$

Benign: if $Pd \leq 0.4$

Measure	Class		
	P	H	B
MCC	-0.12	0.12	-0.25

Some alternative prediction, ex-post...

- What is the molecular effect relevant for the problem at hand?

Could the effect be related to protein stability perturbation?

Thermodynamic data for 141 SAP types found in ProTherm

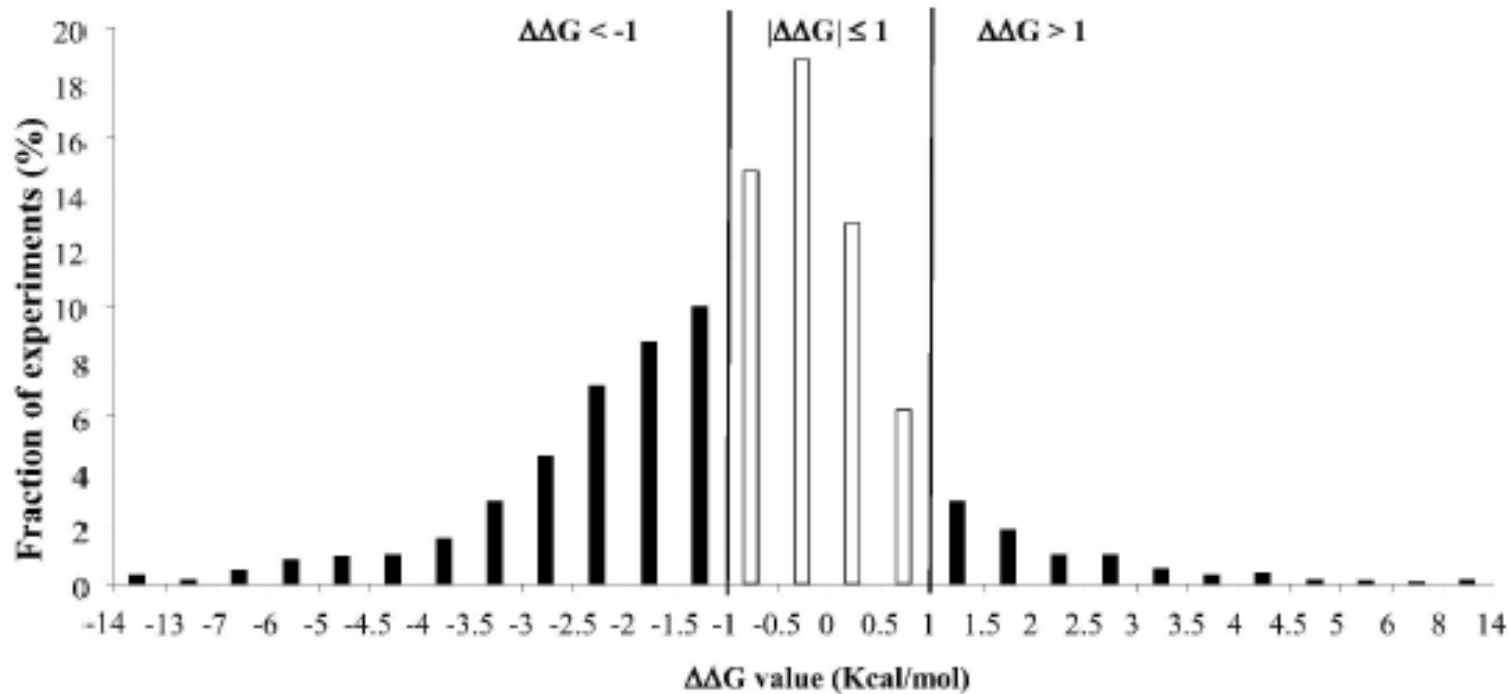
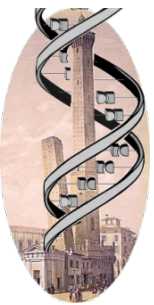


Figure 1. Distribution of $\Delta\Delta G$ values corresponding to 141 SNP types found in ProTherm.

Each bar represents the fraction of experiments (total is 3089) with $\Delta\Delta G$ values in the corresponding range. Mutations corresponding to $\Delta\Delta G < -1$ Kcal/mol and $\Delta\Delta G > 1$ Kcal/mol are labeled as perturbing the protein stability .

ProTherm:

[//gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html](http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html)



Computing the probability indexes of protein incorrect folding and P_p

For each SAP type ($X \rightarrow Y$) we computed the probability index of perturbing protein stability:

$$P_p(X \rightarrow Y) = N_p(X \rightarrow Y) / N(X \rightarrow Y)$$

$N_p(X \rightarrow Y)$ is the number of $X \rightarrow Y$ variations with $|\Delta\Delta G| > 1$ kcal/mol

$N(X \rightarrow Y)$ total number of variations $X \rightarrow Y$ in the thermodynamic database



The Perturbation **Probability Index (Pp)**



	G	A	V	P	L	I	M	F	W	Y	S	T	C	N	Q	H	D	E	K	R
G		.50 70(22)	.70 18(8)						.50* 2(2)		.36 9(7)		.43 5(3)				.67 13(8)	.40* 3(3)		.71 5(4)
A	.58 67(25)		.60 28(12)	.40 13(8)							.27 24(11)	.47 13(9)					.33 4(4)	.33 4(4)		
V	.76 36(16)	.73 122(37)			.46 33(17)	.17 44(22)	.50 14(8)	.64 9(6)									.33 4(2)	.60* 3(2)		
P		.48 60(25)			.57 5(5)								.60* 3(3)		** 0(0)	** 0(0)				.33* 1(1)
L			.67 37(17)	.56 7(6)		.65 24(9)	.23 11(2)	.50 6(4)	.60* 3(3)		.60* 3(3)				.60* 3(3)	.50* 2(2)				.50 6(6)
I			.33 77(29)		.35 24(15)		.60 18(6)	.70 8(7)			.86 5(5)	.94 16(11)		.67* 1(1)					.67* 1(1)	.67* 1(1)
M			.57 5(4)		.29 15(9)	.43 12(8)					.25* 2(2)								.67 4(2)	.67* 1(1)
F			.89 7(7)		.60 23(18)	.67 7(6)				.33 13(11)	.88 6(4)		.75* 2(2)							
W	** 0(0)				.50 4(4)						.67* 1(1)		** 0(0)							.75* 2(2)
Y								.36 54(27)			.80 8(5)		.71 5(5)	.83 4(3)		.80* 3(3)	.88 6(6)			
S	.33 13(10)	.23 45(24)		.40* 3(3)	.17 4(4)	.60* 3(2)		.43 5(3)	.33* 1(1)	.60* 3(3)		.33 10(10)	.71 5(4)	.13 6(5)						.43 5(5)
T		.50 46(23)		.83 4(3)		.44 16(7)	.33* 1(1)				.33 28(14)			.38 6(3)					.33* 1(1)	.33 4(4)
C	.67* 1(1)							** 0(0)	** 0(0)	.67* 1(1)	.81 19(14)									** 0(0)
N					.71 5(4)					.63 6(6)	.25* 2(2)					.50 4(4)	.37 17(11)		.50 6(5)	
Q				.50* 2(2)	.29 5(4)											.25* 2(2)		.17 4(3)	.25* 2(2)	.50* 2(2)
H				.50* 2(2)	.60* 3(3)					.42 10(7)				.67 7(7)	.28 16(9)		.67 4(4)			.50 4(4)
D	.52 19(8)	.47 57(28)	.25* 2(2)							.50* 2(2)				.45 36(21)		.73 9(8)		.18 9(6)		
E	.70 28(11)	.38 69(31)	.64 9(9)												.25 42(21)		.43 12(10)		.33 37(13)	
K						.60* 3(3)	.14 12(7)					.67 4(3)		.57 5(5)	.28 23(7)			.33 25(11)		.21 12(9)
R	.71 15(7)			.33* 1(1)	.40* 3(3)	** 0(0)	.50* 2(2)		.33* 1(1)		.80* 3(3)	.67* 1(1)	.60* 3(3)		.30 8(8)	.56 7(5)			.45 9(6)	

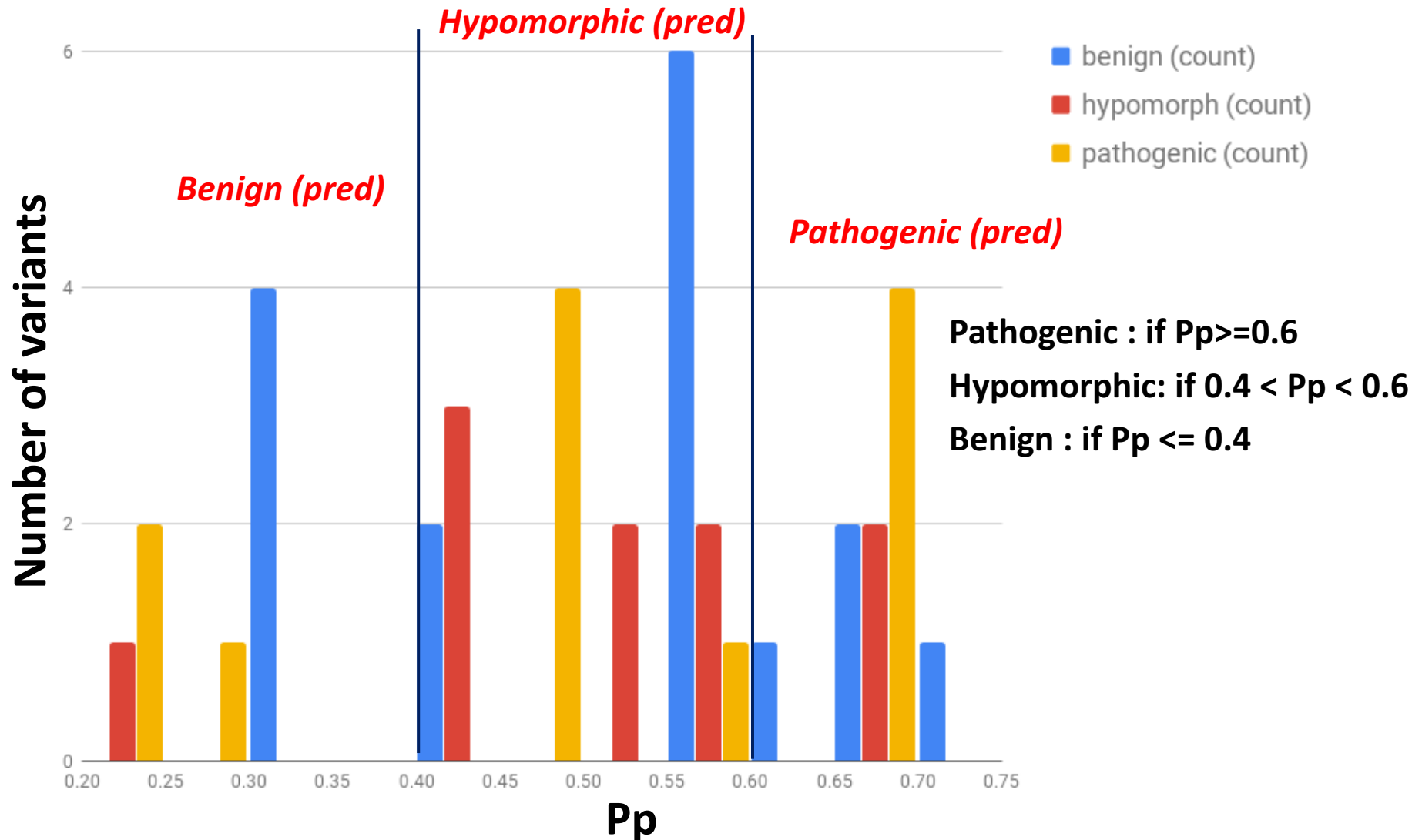
Values of the Perturbation **Probability Index (Pp)** for each SAP type.

The darker is the color, the higher the Pp value.

Wild-type residues are shown on the left, target residues at the top. For each mutation type, the corresponding number of mutations reported in ProTherm and, among round brackets, the number of proteins in which mutations were tested are shown (*labels Pp computed from less than four mutations, 10% of the total). Pp is not computed (**) for a mutation type when data are not available in ProTherm.

Results with Pp matrix

Distribution of Pp values for pathogenic, hypomorphic and benign variants



Results with Pp matrix

		Predicted		
		P	H	B
Observed	P	4	5	3
	H	3	6	1
	B	5	6	5

Pathogenic : if $P_p \geq 0.6$

Hypomorphic: if $0.4 < P_p < 0.6$

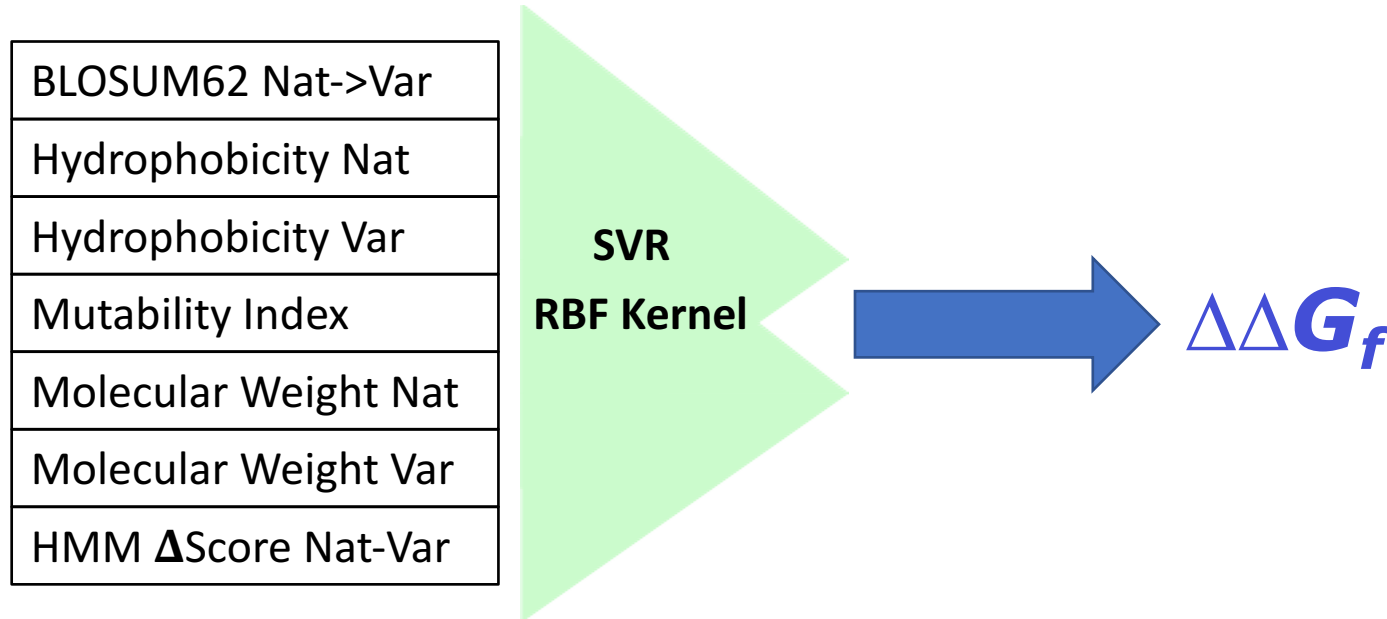
Benign : if $P_p \leq 0.4$

Measure	Class		
	P	H	B
MCC	0.03	0.18	0.15

INPS predictor

- **Impact of Non-synonymous variation on Protein Stability**
(INPS): prediction of $\Delta\Delta G_f$ upon protein single-point variation
- Based on a **Support Vector Regression (SVR)** approach

Fariselli, P., Martelli, P.L., Savojardo, C., Casadio, R. (2015) INPS: predicting the impact of non-synonymous variations on protein stability from sequence. **Bioinformatics**, 31(17), 2816-2821.

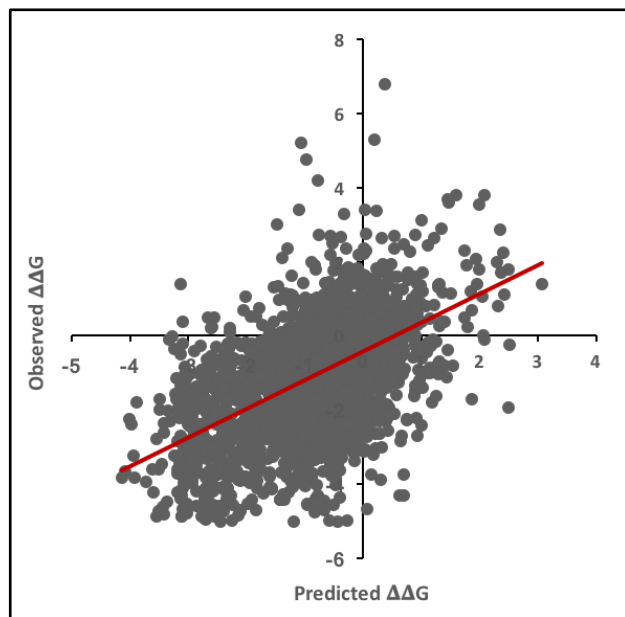


- **Mutability index** is computed for the native residue using the Dayhoff mutability scale
- HMM Δ score
 - Encodes the **evolutionary information**
 - A HMM is built using a MSA of query sequences (native)
 - Both native and variant sequences are aligned to the HMM and the Viterbi-score difference is taken as a **measure of the variation distance between the two sequences**

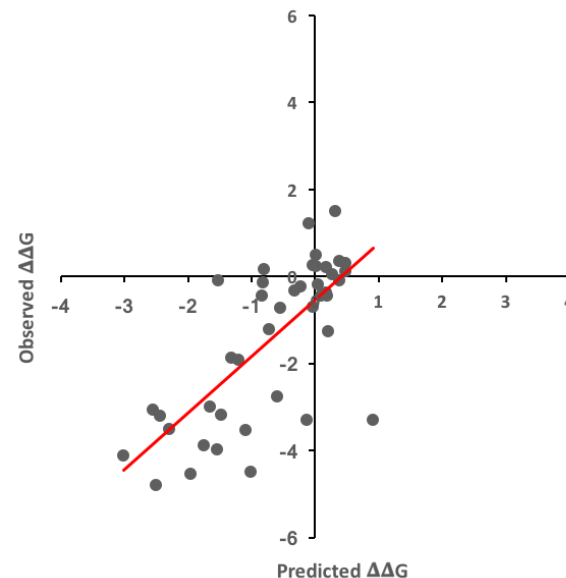
Fariselli, P., Martelli, P.L., Savojardo, C., Casadio, R. (2015) INPS: predicting the impact of non-synonymous variations on protein stability from sequence. **Bioinformatics**, 31(17), 2816-2821.

INPS: performance evaluation

INPS predictions on S2648
(5Fold cross-validation, per protein split)



INPS predictions on P53 dataset



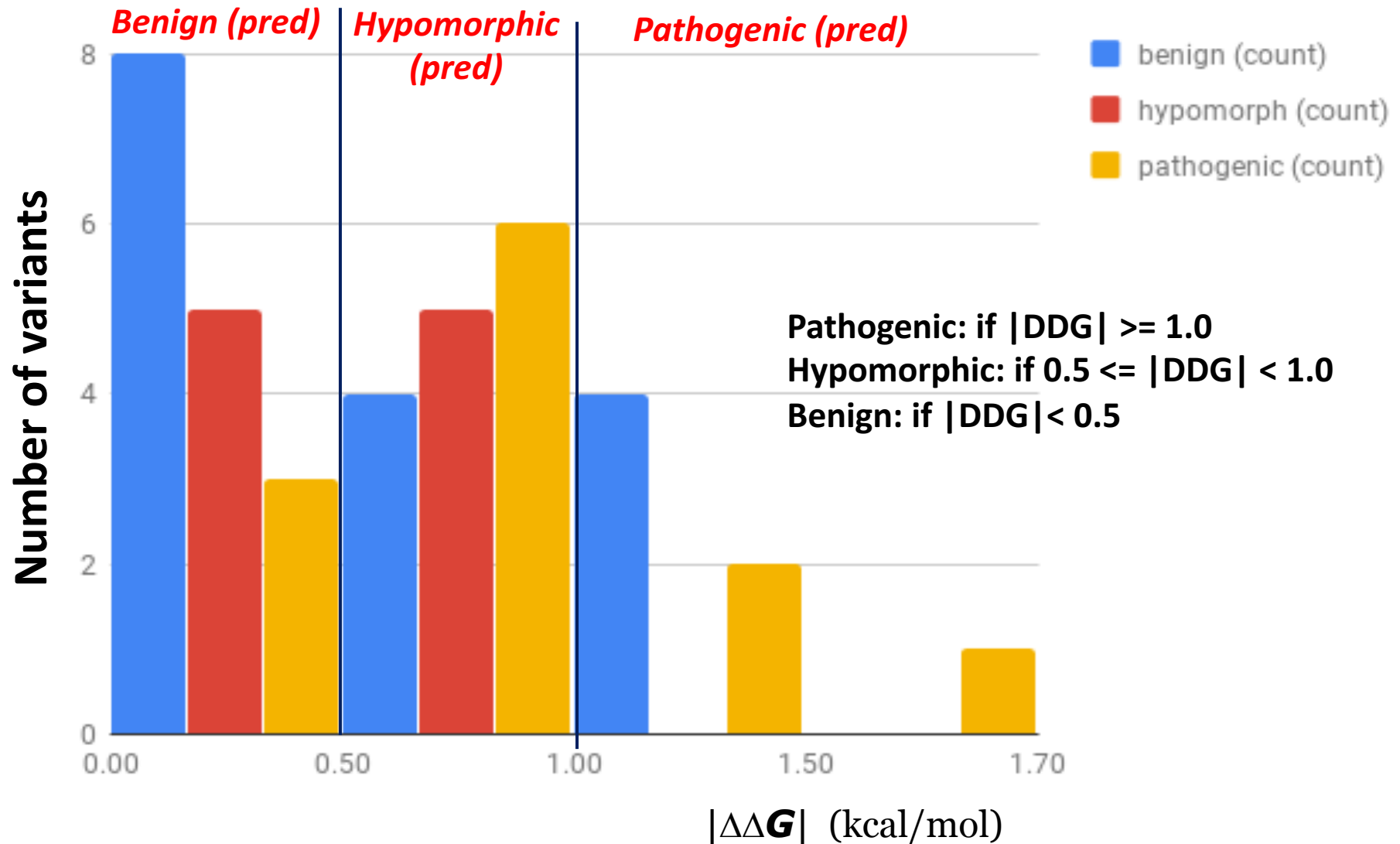
Method	CC	RMSE
INPS	0.53	1.29
mCSM	0.51	1.26

CC= Pearson's correlation coefficient
RMSE=Root Mean Square Error

Method	CC	RMSE
INPS	0.71	1.49
mCSM	0.68	1.40
Duet	0.68	1.39
PopMusic2.0	0.56	1.52

Results with INPS (from sequence)

Distribution of $|\Delta\Delta G|$ values for pathogenic, hypomorphic and benign variants



Results with INPS (from sequence)

		Predicted		
		P	H	B
Observed	P	3	6	3
	H	0	5	5
	B	4	4	8

Pathogenic: if $|DDG| \geq 1.0$
Hypomorphic: if $0.5 \leq |DDG| < 1.0$
Benign: if $|DDG| < 0.5$

Measure	Class		
	P	H	B
MCC	0.12	0.13	0.14



Bologna Biocomputing Group

University of Bologna

[Home](#)[Members](#)[Predictors/Databases](#)[Publications](#)[Training](#)[BWS](#)[Activities](#)[WebMail](#)

Group Leader

Rita Casadio

Rita Casadio



Senior Researchers

Emidio Capriotti
Pietro Di Lena
Piero Fariselli (University of Padova)
Pier Luigi Martelli

Young Researchers

Castrense Savojardo
Giuseppe Profiti (CNR)

External Collaborators

Ivan Rossi (BioDec)

PhD Students

Samuele Bovo
Giulia Babbi

Former Members

Francesco Aggazio
Lisa Bartoli
Remo Calabrese
Raffaele Fronza
Valentina Indio
Paola Marani
Ludovica Montanucci
Andrea Pierleoni
Damiano Piovesan
Deepak Rajan
Priyank Shukla
Shaline Tiwari



Castrense Savojardo



Giulia Babbi



Pier Luigi Martelli

News and Announcement

June 28, 2017

Workshop of the Group of Computational and Systems Biology, Italian Society of Biochemistry and Molecular Biology: Accademia delle Scienze, Bologna, Italy [link](#)

May 29 - June 1, 2017

Advanced course: "Cracking the disease code: mapping genomes to pathogenesis pathways" [Prof. Yana Bromberg](#) -Department of Biochemistry and Microbiology- Rutgers University, New Brunswick, NJ (USA)

April 21, 2017. 15:00-17:30

"The self-consistency test for measuring bias in prediction protein destabilization upon mutation. Finding multi-dimensional epistasis in high-throughput data", [Dmitry Ivankov](#), Center for Genomic Regulation, Barcelona (ES). Lab C, via S. Giacomo 9/2, Bologna

www.biocomp.unibo.it