

# 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.



# CHEK2

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



**Biocomputing Group**  
**University of Bologna**  
**[www.biocomp.unibo.it](http://www.biocomp.unibo.it)**



# Chek2 challenge

SubmissionIDs: 838,839

---

- **Input:** 34 missense variants within the Cell-cycle-checkpoint Kinase (CHEK2) from targeted resequencing of CHEK2 from ~ 1000 Latina breast cancer cases and 1000 ancestry matched controls
- **Challenge:** estimate the probability that a variant occurs in a case (cancer) individual
- **Our approach:**
  - Submission1 (ID:838): **SNPs&GO using UniproKB protein entry O96017**
  - Submission2 (ID:839): **Disease Index Matrix (Casadio *et al.*, 2011)**

# 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<br>91(73)   | .82<br>239(134) |                 |                 |                 |                 |                 | .78<br>33(30)   |                 | .66<br>334(221) |                 | .81<br>100(59)  |                 |                 |                | .82<br>286(144) | .77<br>232(142) |                 | .76<br>64(37)    |
| A | .47<br>58(54)   |                 | .45<br>383(325) | .83<br>39(113)  |                 |                 |                 |                 |                 |                 | .33<br>119(113) | .43<br>415(344) |                 |                 |                 |                | .57<br>181(91)  | .51<br>65(58)   |                 |                  |
| V | .62<br>75(69)   | .33<br>178(162) |                 |                 | .42<br>117(105) | .16<br>248(239) | .50<br>223(185) | .69<br>54(47)   |                 |                 |                 |                 |                 |                 |                 |                | .83<br>44(38)   | .83<br>31(25)   |                 |                  |
| P |                 | .43<br>68(67)   |                 |                 | .47<br>443(358) |                 |                 |                 |                 |                 | .48<br>224(184) | .55<br>84(74)   |                 |                 | .40*<br>38(38)  | .51*<br>47(42) |                 |                 |                 | .73<br>92(75)    |
| L |                 |                 | .43<br>114(101) | .77<br>473(295) |                 | .14<br>46(45)   | .21<br>66(62)   | .50<br>157(157) | .57<br>10(10)   |                 | .56<br>71(60)   |                 |                 |                 | .77<br>40(37)   | .63<br>25(22)  |                 |                 |                 | .86<br>134(106)  |
| I |                 |                 | .16<br>191(185) |                 | .33<br>38(37)   |                 | .47<br>84(80)   | .87<br>74(20)   |                 |                 | .72<br>42(37)   | .62<br>262(154) | .90<br>56(49)   |                 |                 |                |                 |                 | .78<br>6(5)     | .65<br>13(12)    |
| M |                 |                 | .47<br>146(123) |                 | .34<br>43(42)   | .45<br>87(79)   |                 |                 |                 |                 |                 | .41<br>143(128) |                 |                 |                 |                |                 |                 | .73<br>29(28)   | .78<br>41(39)    |
| F |                 |                 |                 | .78<br>33(29)   |                 | .59<br>137(111) | .62<br>25(24)   |                 |                 | .31<br>12(12)   | .81<br>50(83)   |                 | .76<br>47(41)   |                 |                 |                |                 |                 |                 |                  |
| W | .74*<br>38(33)  |                 |                 |                 |                 | .75<br>10(15)   |                 |                 |                 |                 | .88<br>22(21)   |                 | .92*<br>61(46)  |                 |                 |                |                 |                 |                 | .66<br>130(118)  |
| Y |                 |                 |                 |                 |                 |                 |                 | .41<br>10(10)   |                 |                 | .77<br>31(26)   | .80<br>246(166) | .75<br>28(26)   | .62<br>91(84)   | .81<br>45(37)   |                |                 |                 |                 |                  |
| S | .34<br>78(74)   | .19<br>58(50)   | .55<br>168(147) | .56<br>138(126) | .39<br>62(57)   |                 | .45<br>149(154) | .55<br>7(7)     | .51<br>43(41)   |                 | .24<br>78(99)   | .47<br>64(63)   | .28<br>163(158) |                 |                 |                |                 |                 |                 | .62<br>145(90)   |
| T |                 | .27<br>289(199) | .58<br>75(72)   |                 | .50<br>215(182) | .49<br>178(168) |                 |                 |                 |                 | .22<br>49(67)   |                 | .42<br>55(51)   |                 |                 |                |                 |                 | .52<br>38(34)   | .67<br>45(41)    |
| C | .78<br>41(39)   |                 |                 |                 |                 |                 | .85*<br>82(59)  | .82*<br>51(35)  | .83<br>266(125) | .74<br>78(51)   |                 |                 |                 |                 |                 |                |                 |                 |                 | .75*<br>208(140) |
| N |                 |                 |                 |                 |                 | .84<br>34(29)   |                 |                 |                 | .72*<br>28(28)  | .49<br>218(189) | .56<br>52(51)   |                 |                 |                 | .49<br>35(35)  | .46<br>81(77)   |                 | .59<br>121(107) |                  |
| Q |                 |                 | .75<br>45(41)   | .62<br>25(25)   |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 | .37<br>117(105) |                | .44<br>55(52)   | .43<br>84(87)   | .37<br>186(172) |                  |
| H |                 |                 | .68<br>58(51)   | .76<br>28(25)   |                 |                 |                 |                 |                 | .58<br>113(91)  |                 |                 | .54<br>23(22)   | .45<br>49(44)   |                 | .69<br>35(35)  |                 |                 | .50<br>181(145) |                  |
| D | .66<br>147(125) | .59<br>40(36)   | .79<br>81(66)   |                 |                 |                 |                 |                 |                 | .73<br>103(79)  |                 |                 | .52<br>281(238) |                 | .69<br>74(64)   |                | .37<br>112(110) |                 |                 |                  |
| E | .53<br>133(114) | .51<br>84(80)   | .67<br>44(39)   |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 |                 | .45<br>86(75)   |                | .35<br>123(116) |                 | .48<br>47(33)   | .61<br>145(118)  |
| K |                 |                 |                 |                 |                 | .68<br>18(17)   | .70<br>6(6)     |                 |                 |                 | .52<br>36(35)   | .51<br>114(99)  | .43<br>49(47)   |                 |                 |                | .48<br>157(142) |                 |                 | .32<br>118(115)  |
| R | .68<br>286(167) |                 | .84<br>74(135)  | .77<br>127(101) | .47*<br>15(15)  | .21<br>15(15)   |                 | .76<br>412(248) |                 | .60<br>122(108) | .71<br>44(41)   | .75<br>495(328) | .59<br>318(298) | .64<br>472(358) |                 |                |                 |                 |                 | .21<br>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.

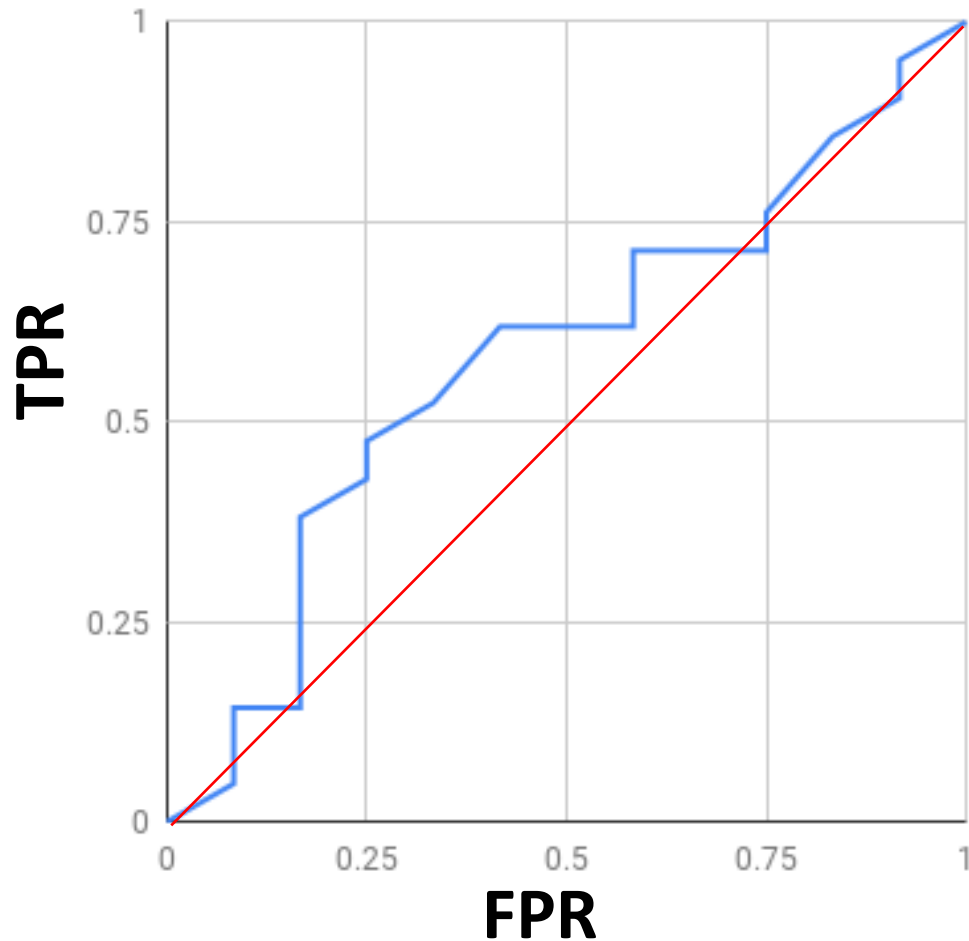
For each mutation type the corresponding number of mutations and, between round brackets, the number of proteins in which mutations are annotated, are shown. Pd is computed for the 150 SAP types due to SNPs. Squares comprising mutations within the same residue group are coloured in red, blue, green and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

# Pd: Calibration

- |              |                        |            |
|--------------|------------------------|------------|
| • $Pd < 0.4$ | $P(\text{case}) = 0.5$ | NEUTRAL    |
| • $Pd > 0.6$ | $P(\text{case}) = 1$   | PATHOGENIC |

We did not considered the possibility to predict BENIGN or PROTECTIVE variation

# Rehabilitating Pd...



**AUC = 0.58**

**MCC = 0.22**

**[with  $P(\text{case}) = 0.8$ ]**

# Functional Annotations Improve the Predictive Score of Human Disease-Related Mutations in Proteins

Remo Calabrese, Emidio Capriotti, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio\*

*Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy*

Communicated by Mauno Vihinen

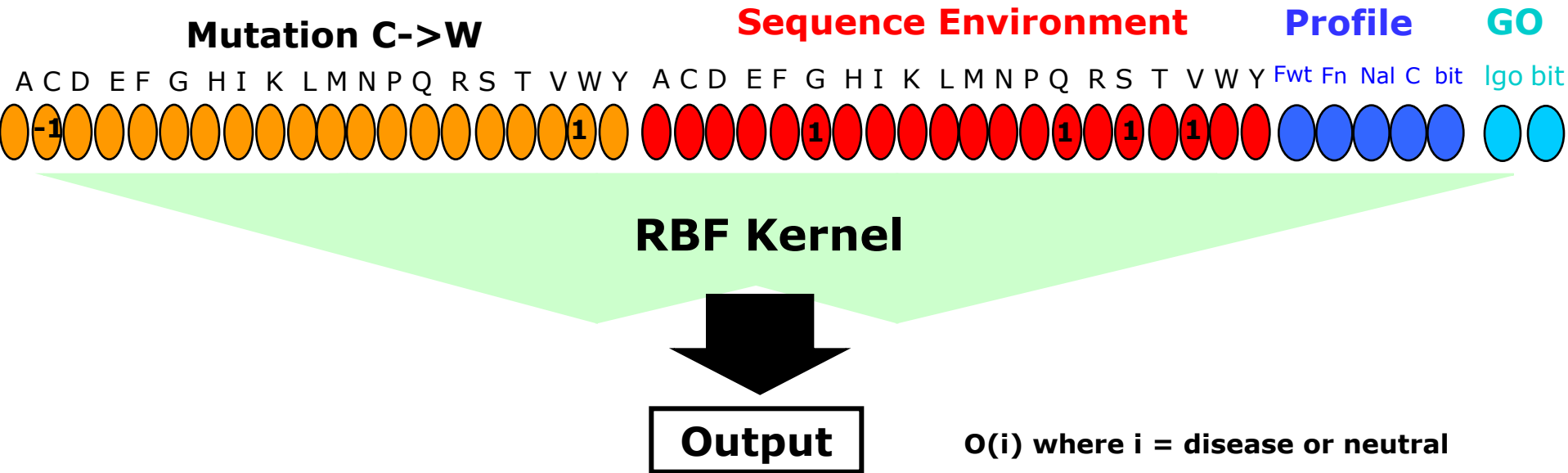
Received 16 December 2008; accepted revised manuscript 1 May 2009.

Published online 12 May 2009 in Wiley InterScience ([www.interscience.wiley.com](http://www.interscience.wiley.com)). DOI 10.1002/humu.21047

## The role of function/SNPs&GO:

an accurate method that, starting from a protein sequence, can predict whether a mutation is disease related or not by exploiting the protein functional annotation

# SNPs&GO: the SEQUENCE-PROFILE AND GO-BASED PREDICTOR



**Svm-Seq-Prof-GO: Mutation+ Sequence Environment + Profile + GO**

**Fwt, Fn=** occurrence of wild-type and mutated residue in sequence Profile

**Nal=** Number of sequences aligned at the mutated position

**C=** Conservation index based on variance at the mutated position

**bit=** flag for the presence/absence of the above information

**LGO=** log-odd ratio of the frequency with which any GO term is found in Disease and Neutral sets

**bit=** flag for the presence/absence of the above information:

$$LGO = \sum \log_2 [(GTO(GO)_D + 1) / (GTO(GO)_N + 1)]$$

**Table 3. Bench Marking of SNPs&GO with Other Predictive Methods**

| Method                  | Q2   | P[D] | Q[D] | P[N] | Q[N] | C    | PM (%) |
|-------------------------|------|------|------|------|------|------|--------|
| PolyPhen <sup>a</sup>   | 0.71 | 0.76 | 0.75 | 0.63 | 0.64 | 0.39 | 58     |
| SIFT <sup>b</sup>       | 0.76 | 0.75 | 0.76 | 0.77 | 0.75 | 0.52 | 93     |
| PANTHER <sup>c</sup>    | 0.74 | 0.77 | 0.73 | 0.71 | 0.76 | 0.48 | 76     |
| Eremorph <sup>d</sup>   | 0.74 | 0.83 | 0.64 | 0.68 | 0.85 | 0.50 | 82     |
| HybridMeth <sup>e</sup> | 0.74 | 0.74 | 0.70 | 0.74 | 0.77 | 0.47 | 100    |
| SNPs&GO                 | 0.82 | 0.83 | 0.78 | 0.80 | 0.85 | 0.63 | 100    |

<sup>a</sup>Downloaded from the Web server <http://genetics.bwh.harvard.edu/pph/data/index.html>.

<sup>b</sup>Downloaded from <http://blocks.fhcrc.org/sift/SIFT.html>.

<sup>c</sup>Downloaded from <http://www.pantherdb.org/downloads/index.jsp>.

<sup>d</sup>Our previous implementation [Capriotti et al., 2006]. Performances are computed by training/testing on the SP\_human data set (see Materials and Methods). Only HybridMeth and SNPs&GO are scored with a real crossvalidation procedure, because all the others were trained on sets that were possibly homologous to ours. PM is the percentage of predicted mutations. For the definition of the scoring indexes see the Materials and Methods section.

<sup>e</sup>Downloaded from the web server <http://discovery.swmed.edu/eremorph/>. Downloadable (a and c) predictors were run locally; otherwise, data are retrieved via Web server queries (a and d).

**Table 3. Performance of Prediction Methods**

|                                                                                                 | MutPred | nsSNPAnalyzer | Panther | PhD-SNP | PolyPhen1a | PolyPhen 1b | PolyPhen 2a | PolyPhen 2b | SIFT  | SNAP  | SNPs&GO |
|-------------------------------------------------------------------------------------------------|---------|---------------|---------|---------|------------|-------------|-------------|-------------|-------|-------|---------|
| <i>Performance of prediction methods (full data)</i>                                            |         |               |         |         |            |             |             |             |       |       |         |
| tp <sup>a</sup>                                                                                 | 13829   | 4360          | 9689    | 11900   | 10093      | 14285       | 13807       | 16206       | 10464 | 16000 | 13736   |
| fn <sup>a</sup>                                                                                 | 2507    | 2778          | 2859    | 6896    | 9185       | 4993        | 5102        | 2703        | 4856  | 2146  | 5487    |
| tn <sup>a</sup>                                                                                 | 15891   | 1319          | 8676    | 16788   | 17669      | 13671       | 13863       | 10199       | 12188 | 8190  | 17028   |
| fp <sup>a</sup>                                                                                 | 4557    | 943           | 2797    | 4377    | 3199       | 7197        | 6010        | 9674        | 7433  | 6387  | 1382    |
| cases + <sup>a</sup>                                                                            | 16336   | 7138          | 12548   | 18796   | 19278      | 19278       | 18909       | 18909       | 15320 | 18146 | 19223   |
| cases - <sup>a</sup>                                                                            | 20448   | 2262          | 11473   | 21165   | 20868      | 20868       | 19873       | 19873       | 19621 | 14577 | 18410   |
| Accuracy <sup>b</sup>                                                                           | 0.81    | 0.60          | 0.76    | 0.71    | 0.69       | 0.70        | 0.71        | 0.69        | 0.65  | 0.72  | 0.82    |
| Precision <sup>b</sup>                                                                          | 0.79    | 0.59          | 0.76    | 0.75    | 0.77       | 0.68        | 0.71        | 0.64        | 0.64  | 0.67  | 0.90    |
| Specificity <sup>b</sup>                                                                        | 0.78    | 0.58          | 0.76    | 0.79    | 0.85       | 0.66        | 0.70        | 0.51        | 0.62  | 0.56  | 0.92    |
| Sensitivity <sup>b</sup>                                                                        | 0.85    | 0.61          | 0.77    | 0.63    | 0.52       | 0.74        | 0.73        | 0.86        | 0.68  | 0.88  | 0.71    |
| NPV <sup>b</sup>                                                                                | 0.84    | 0.60          | 0.77    | 0.68    | 0.64       | 0.72        | 0.72        | 0.78        | 0.66  | 0.83  | 0.76    |
| MCC <sup>b</sup>                                                                                | 0.63    | 0.19          | 0.53    | 0.43    | 0.39       | 0.40        | 0.43        | 0.39        | 0.30  | 0.47  | 0.65    |
| <i>Performance of prediction methods (3D structure)</i>                                         |         |               |         |         |            |             |             |             |       |       |         |
| tp <sup>a</sup>                                                                                 | 5625    | 2857          | 3934    | 5041    | 4563       | 5980        | 5814        | 6726        | 4303  | 6751  | 5887    |
| fn <sup>a</sup>                                                                                 | 517     | 1603          | 1009    | 2411    | 3074       | 1657        | 1842        | 930         | 1329  | 714   | 1746    |
| tn <sup>a</sup>                                                                                 | 1101    | 569           | 735     | 1090    | 1361       | 1070        | 1163        | 843         | 904   | 700   | 1378    |
| fp <sup>a</sup>                                                                                 | 697     | 527           | 441     | 754     | 462        | 753         | 672         | 992         | 901   | 777   | 318     |
| cases + <sup>a</sup>                                                                            | 6142    | 4460          | 4943    | 7452    | 7637       | 7637        | 7656        | 7656        | 5632  | 7465  | 7633    |
| cases - <sup>a</sup>                                                                            | 1798    | 1096          | 1176    | 1844    | 1823       | 1823        | 1835        | 1835        | 1805  | 1477  | 1696    |
| Accuracy <sup>b</sup>                                                                           | 0.76    | 0.58          | 0.71    | 0.63    | 0.67       | 0.68        | 0.70        | 0.67        | 0.63  | 0.69  | 0.79    |
| Precision <sup>b</sup>                                                                          | 0.70    | 0.57          | 0.68    | 0.62    | 0.70       | 0.65        | 0.67        | 0.62        | 0.60  | 0.63  | 0.80    |
| Specificity <sup>b</sup>                                                                        | 0.61    | 0.52          | 0.63    | 0.59    | 0.75       | 0.59        | 0.63        | 0.46        | 0.50  | 0.47  | 0.81    |
| Sensitivity <sup>b</sup>                                                                        | 0.92    | 0.64          | 0.80    | 0.68    | 0.60       | 0.78        | 0.76        | 0.88        | 0.76  | 0.90  | 0.77    |
| NPV <sup>b</sup>                                                                                | 0.88    | 0.59          | 0.75    | 0.65    | 0.65       | 0.73        | 0.72        | 0.79        | 0.68  | 0.83  | 0.78    |
| MCC <sup>b</sup>                                                                                | 0.55    | 0.16          | 0.43    | 0.27    | 0.35       | 0.38        | 0.40        | 0.37        | 0.27  | 0.42  | 0.58    |
| <i>Performance of prediction methods (pathogenic dataset only from LSDBs, not in SwissProt)</i> |         |               |         |         |            |             |             |             |       |       |         |
| tp                                                                                              | 2240    | 1175          | 1368    | 1436    | 1651       | 2410        | 2190        | 2764        | 2131  | 2615  | 2547    |
| fn                                                                                              | 899     | 862           | 1252    | 2158    | 1943       | 1184        | 1361        | 787         | 1145  | 917   | 952     |
| tn                                                                                              | 2655    | 212           | 1508    | 2842    | 3004       | 2333        | 2334        | 1705        | 2073  | 1382  | 2898    |
| fp                                                                                              | 804     | 165           | 501     | 752     | 534        | 1205        | 1028        | 1657        | 1268  | 1069  | 259     |
| cases + <sup>a</sup>                                                                            | 3139    | 2037          | 2620    | 3594    | 3594       | 3594        | 3551        | 3551        | 3276  | 3532  | 3499    |
| cases - <sup>a</sup>                                                                            | 3459    | 377           | 2009    | 3594    | 3538       | 3538        | 3362        | 3362        | 3341  | 2451  | 3157    |
| Accuracy <sup>b</sup>                                                                           | 0.74    | 0.57          | 0.64    | 0.6     | 0.65       | 0.66        | 0.66        | 0.64        | 0.64  | 0.65  | 0.82    |
| Precision <sup>b</sup>                                                                          | 0.75    | 0.57          | 0.68    | 0.66    | 0.75       | 0.66        | 0.67        | 0.61        | 0.63  | 0.63  | 0.90    |
| Specificity <sup>b</sup>                                                                        | 0.77    | 0.56          | 0.75    | 0.79    | 0.85       | 0.66        | 0.69        | 0.51        | 0.62  | 0.56  | 0.92    |
| Sensitivity <sup>b</sup>                                                                        | 0.71    | 0.58          | 0.52    | 0.4     | 0.46       | 0.67        | 0.62        | 0.78        | 0.65  | 0.74  | 0.73    |
| NPV <sup>b</sup>                                                                                | 0.73    | 0.57          | 0.61    | 0.57    | 0.61       | 0.67        | 0.64        | 0.70        | 0.64  | 0.68  | 0.77    |
| MCC <sup>b</sup>                                                                                | 0.48    | 0.14          | 0.28    | 0.21    | 0.33       | 0.33        | 0.31        | 0.30        | 0.27  | 0.31  | 0.66    |

<sup>a</sup>Total number of cases used by the given program (not normalized).

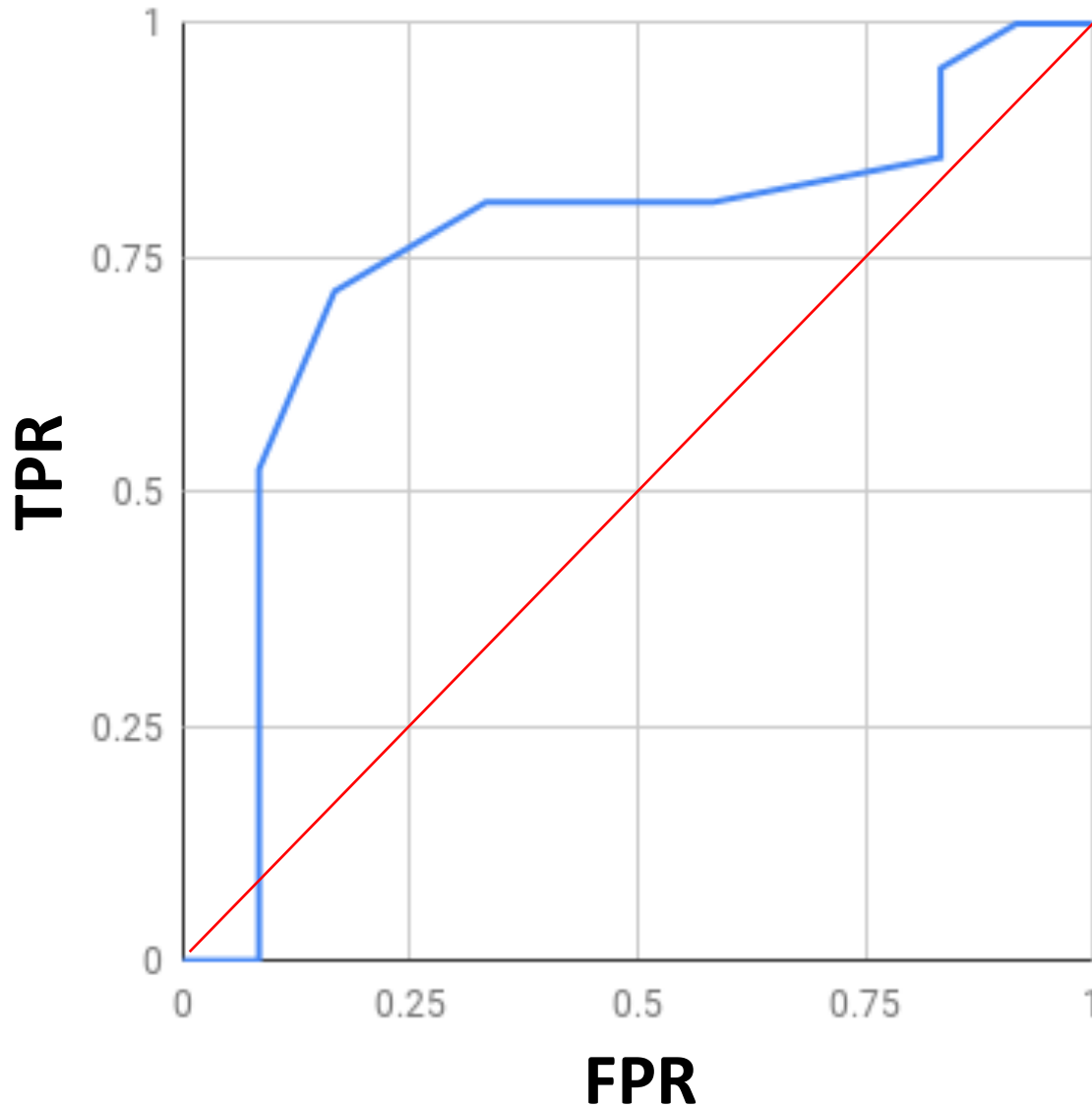
<sup>b</sup>Accuracy, precision, specificity, sensitivity, NPV, and MCC are calculated from normalised numbers.

# SNPs&GO: Calibration

- Based on prediction and Reliability Index
  - Neutral ,  $RI = 10 \rightarrow P = 0.5$  NEUTRAL
  - Neutral,  $RI = 0 \rightarrow P = 0.75$
  - Disease,  $RI = 0 \rightarrow P = 0.75$
  - Disease,  $RI = 10 \rightarrow P = 1$  PATHOGENIC

We did not considered the possibility to predict BENIGN or PROTECTIVE variation

# ROC



**AUC = 0.75**

# As a binary prediction

## Prediction

Data

|                | Pathogenic | Neutral |
|----------------|------------|---------|
| Pathogenic     | 15         | 6       |
| Neutral/Benign | 2          | 10      |

**MCC = 0.53**

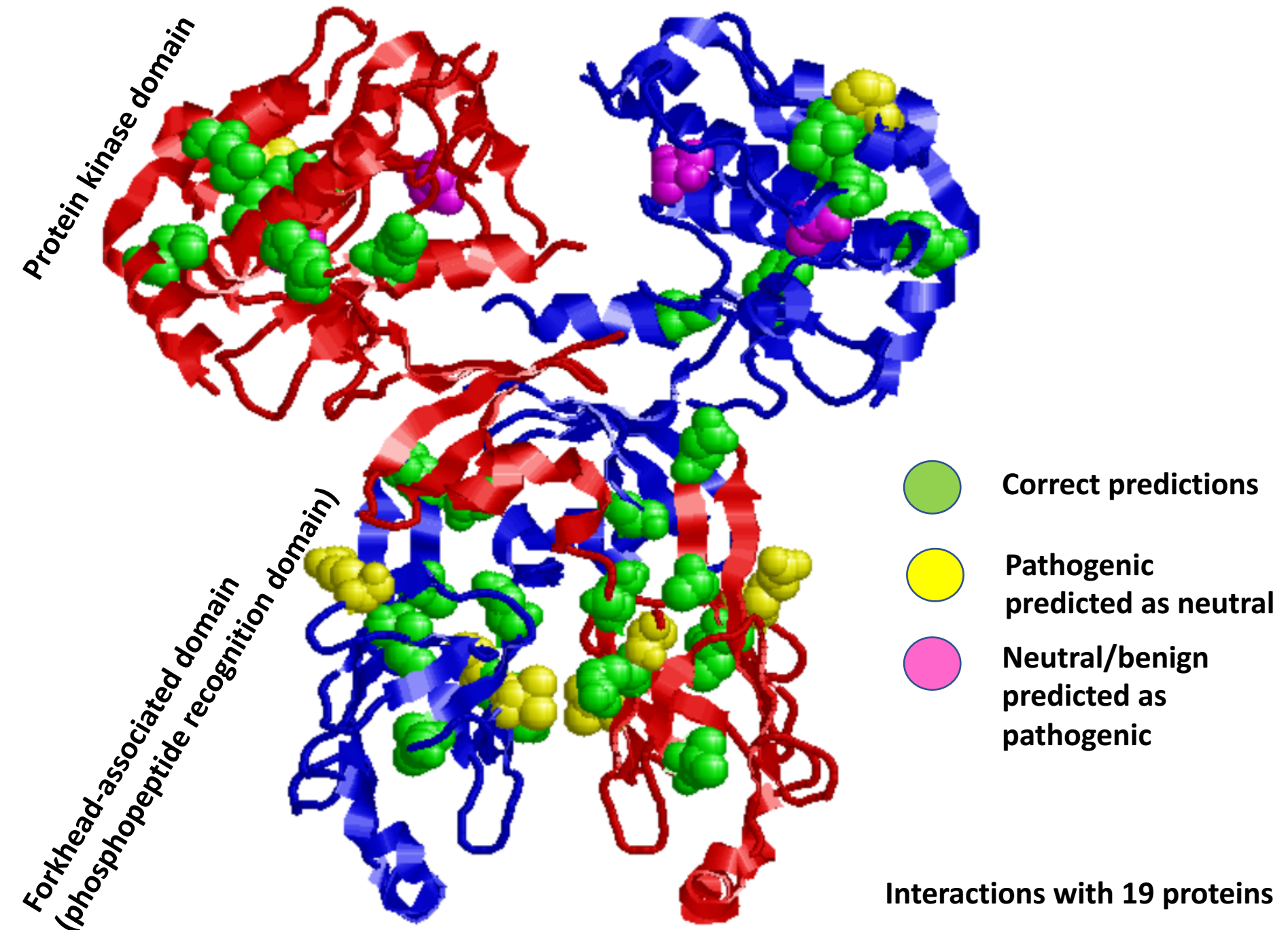
**Recall (Neutral) = 0.83**

**Precision (Neutral) = 0.63**

**Accuracy = 0.74**

**Recall (Pathogenic) = 0.71**

**Precision (Pathogenic) = 0.88**





# 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](http://www.biocomp.unibo.it)