

CAGI slide pack guidelines

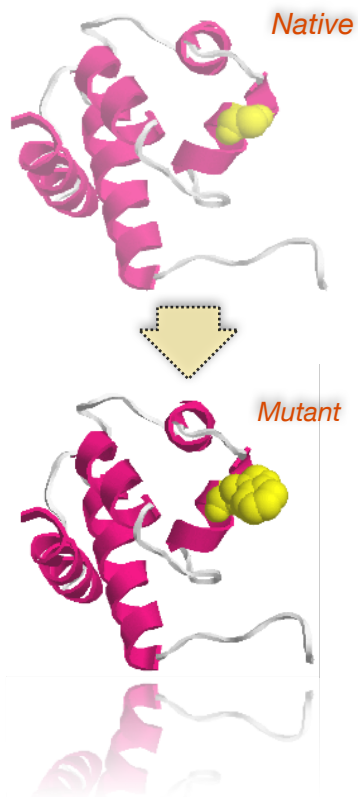
I agree to my slides being used according to the following terms:

Emidio Capriotti

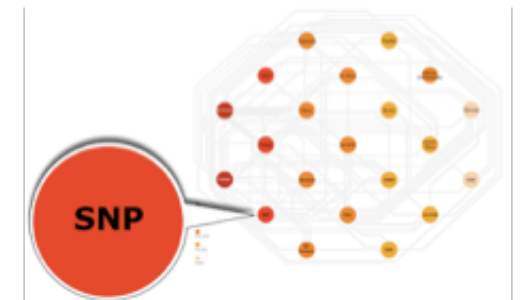
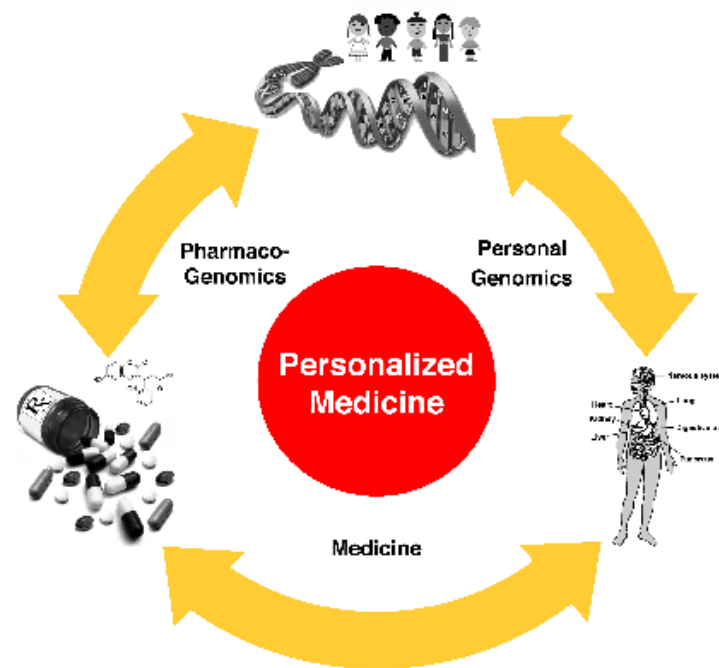
- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix 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.



Predicting the effect of variants on GAA function



CAGI 5 Conference
Chicago (IL), July 6, 2018



Emidio Capriotti
<http://biofold.org/>



**Biomolecules
Folding and
Disease**

Department of Pharmacy
and Biotechnology (FaBiT)
University of Bologna



GAA Challenge

- Participants were asked to submit predictions on the effect of the variants on GAA enzymatic activity.
- Prediction ranges from 0 (no activity) to 1 (wild-type level of activity), or >1 if the predicted activity is greater than wild-type activity.
- The effect of 357 mutations has been experimentally determined.
- ~300 mutation maps on protein structure 5KZW



- Trefoil 82-130 (Green)
- NtCtMGAM_N 147-253 (Red)
- Gal_mutarotas_2 254-320 (Blue)
- Glyco_hydro_31 340-824 (Yellow)

Submitted predictions

For the GAA challenge we submitted predictions using the following methods:

- **PhD-SNP**: protein sequence-based features
- **SNPs&GO**: protein sequence-based features and functional information
- **PhD-SNPg**: sequence-based algorithm using nucleotide sequence
- **Meta-SNP**: meta predictor using PhD-SNP, PANTHER, SIFT and SNAP

The prediction represents the probabilistic score of deleterious variant subtracted from one.

I **evaluated structure-based** predictions (PhD-SNP^{3D}) **not submitted to the challenge**

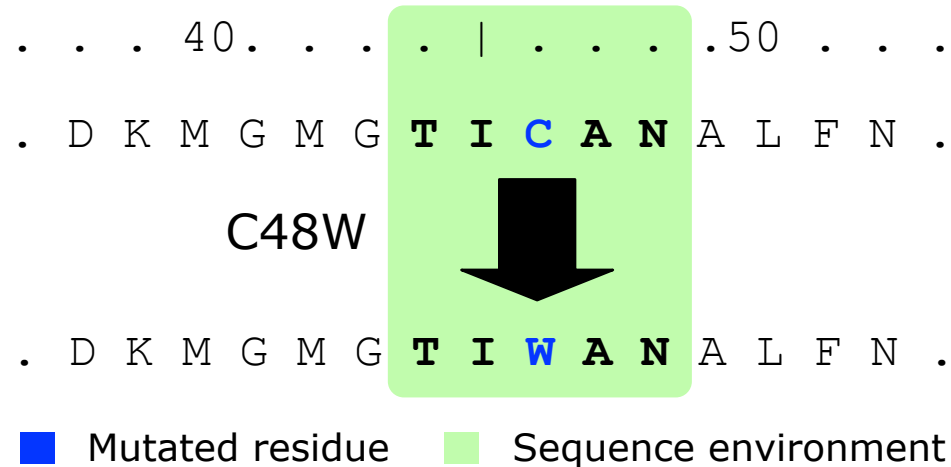
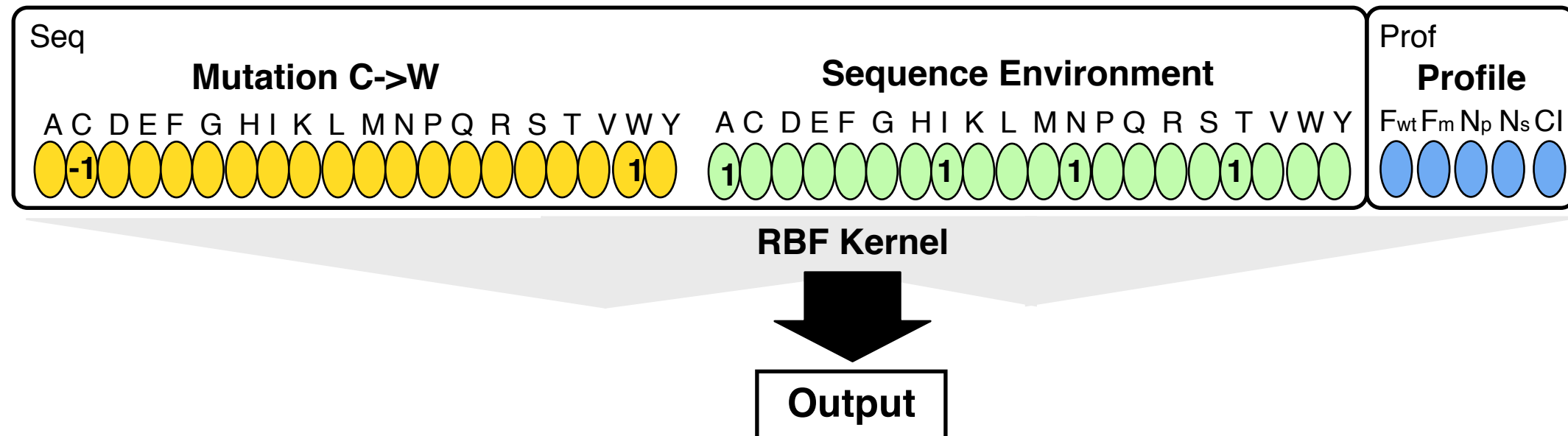
Evaluation measures

All predictions have been evaluated calculating

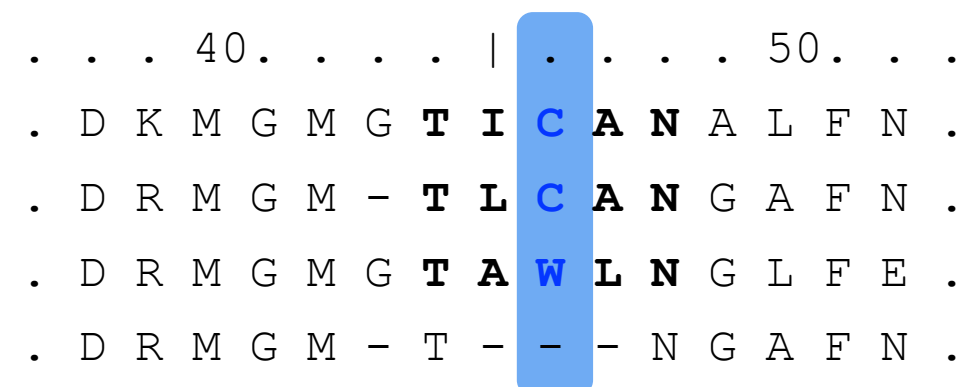
- Pearson Correlation Coefficient
- Spearman Rank Correlation Coefficient
- Kendall Tau Rank Correlation Coefficient
- RMSE (Root Mean Square Error)
- MAE (Mean Absolute Error)

Predictions are also evaluated casting the problem to a **binary classification task**.

PhD-SNP method



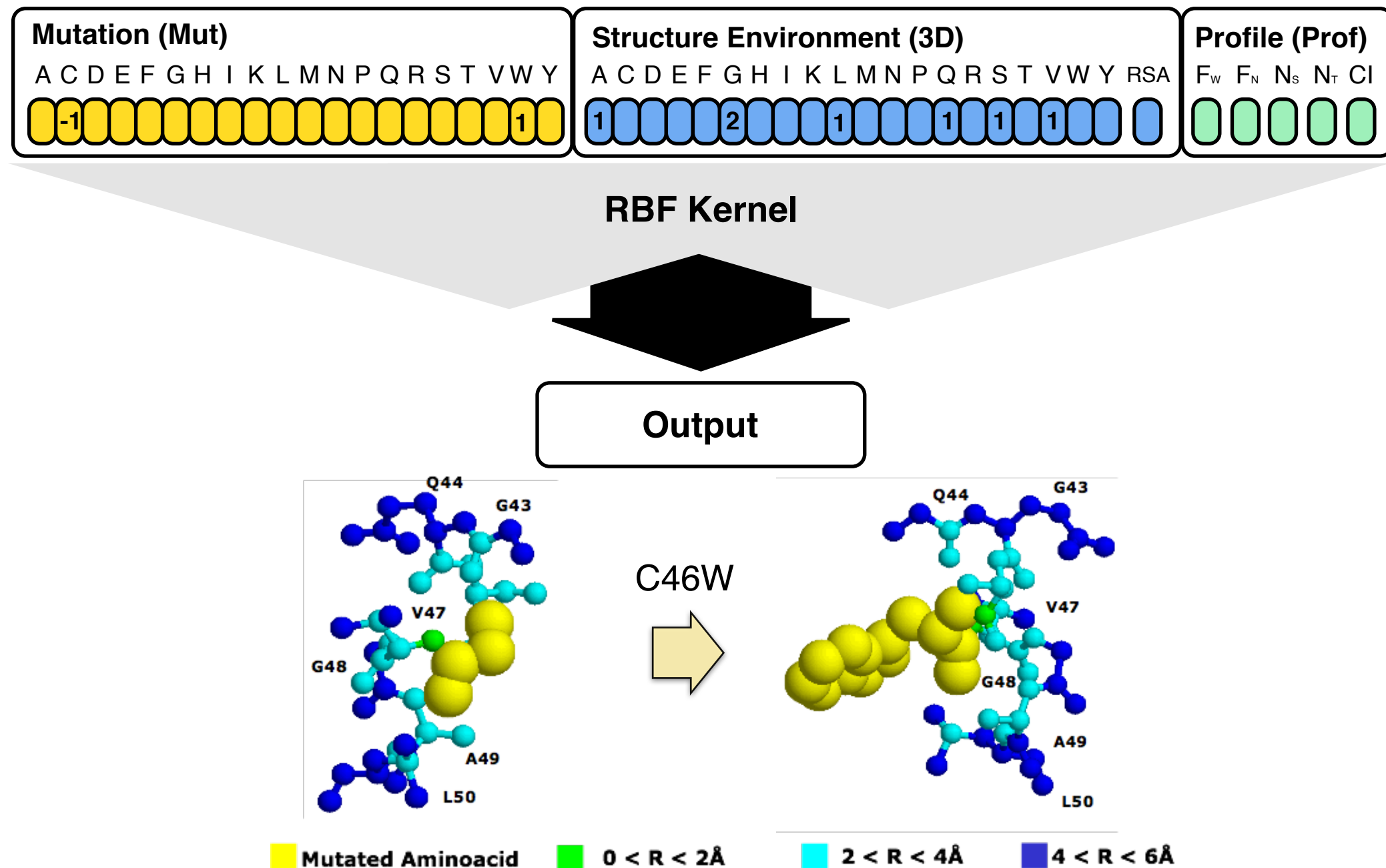
Sequence information is encoded in 2 vectors each one composed by 20 elements. The first vector encodes for the mutation and the second one for the sequence environment



Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features

The structure-based method

The method takes in to account 3 different types of information encoded in a 46 elements vector. The input features are: mutation data; structure environment, and sequence profile.



Conclusions

- For the GAA challenge simple sequence-based approach performs better than more sophisticated methods. Although the correlation coefficient are significant the RMSE values are still high.
- Prediction performance depends on the quality of multiple sequence alignment. Predictions in sites with more than 200 aligned sequences are more accurate
- Structure-based predictor results in better performance than sequence-based predictor.
- The analysis of predictions falling in the 4 main domains of the protein confirms that the number of aligned sequences is an important factor in the predictions.

Acknowledgments

Structural Genomics @CNAG

Marc A. Marti-Renom

Davide Bau
David Dufour
Francois Serra

Computational Biology and Bioinformatics Research Group (UIB)

Jairo Rocha

Division of Informatics at UAB

Rui Tian
Shivani Viradia
Malay Basu
Diego E. Penha

Helix Group (Stanford University)

Russ B. Altman

Jennifer Lahti
Tianyun Liu
Grace Tang

Bologna Biocomputing Group

Rita Casadio

Pier Luigi Martelli

University of Padova

Piero Fariselli

University of Camerino

Mario Compiani

Mathematical Modeling of Biological Systems (University of Düsseldorf)

Markus Kollmann

Other Collaborations

Yana Bromberg, Rutgers University, NJ
Francisco Melo, Universidad Catolica, Chile
Sean Mooney, University of Washington, WA
Cedric Notredame, CRG Barcelona
Gustavo Parisi, Universidad de Quilmes
Frederic Rousseau, KU Leuven
Joost Schymkowitz, KU Leuven

FUNDING

Startup funding Dept. of Pathology UAB
NIH:3R00HL111322-04S1 Co-Investigator
EMBO Short Term Fellowship
Marie Curie International Outgoing Grant
Marie Curie Reintegration Grant
Marco Polo Research Project
BIOSAPIENS Network of Excellence
SPINNER Consortium

Biomolecules, Folding and Disease



<http://biofold.org/>