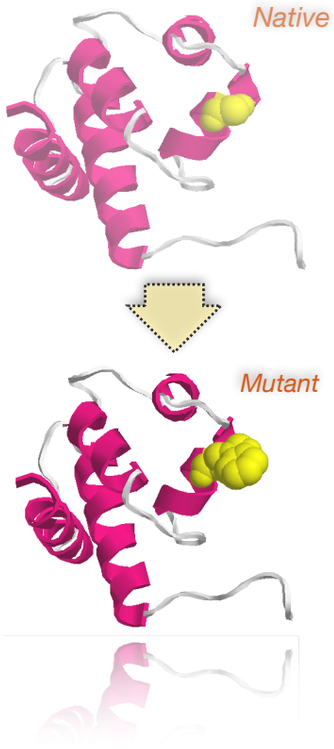
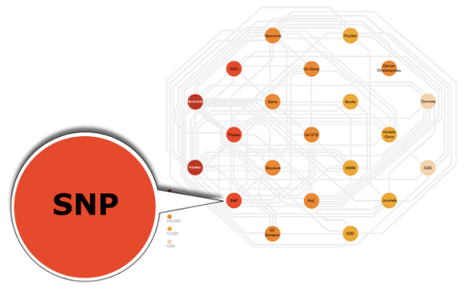
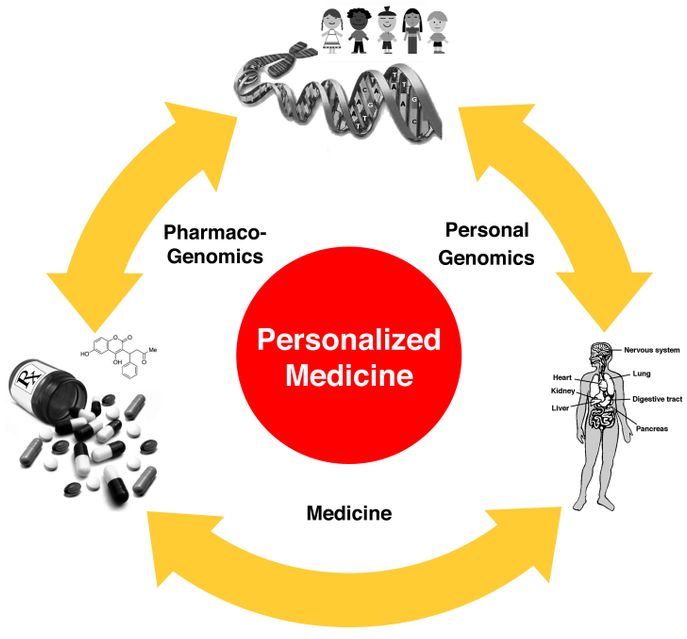
Predicting the impact of genetic variants on protein stability and human health



Winter School - University of Verona
Canazei (TN)
January 17, 2019





OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Department of Pharmacy
and Biotechnology (FaBiT)
University of Bologna

Single Nucleotide Variants

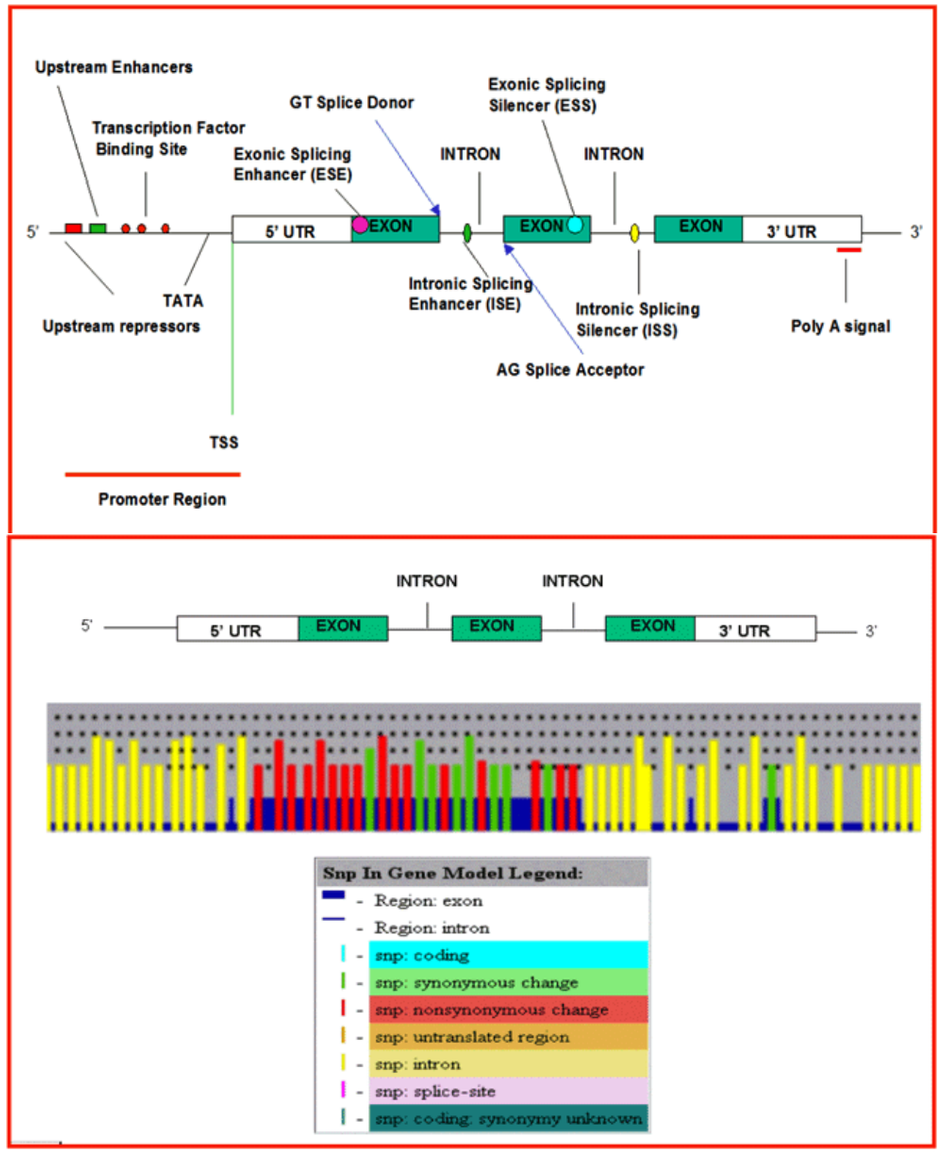
Single Nucleotide Variants (SNVs)
is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome differs between members of the species.
It is used to refer to Polymorphisms when the population frequency is ≥ 1%

SNVs occur at any position and can be classified on the base of their locations.

Coding SNVs can be subdivided into two groups:

Synonymous: when single base substitutions do not cause a change in the resultant amino acid

Non-synonymous or Single Amino Acid Variants (SAVs): when single base substitutions cause a change in the resultant amino acid.

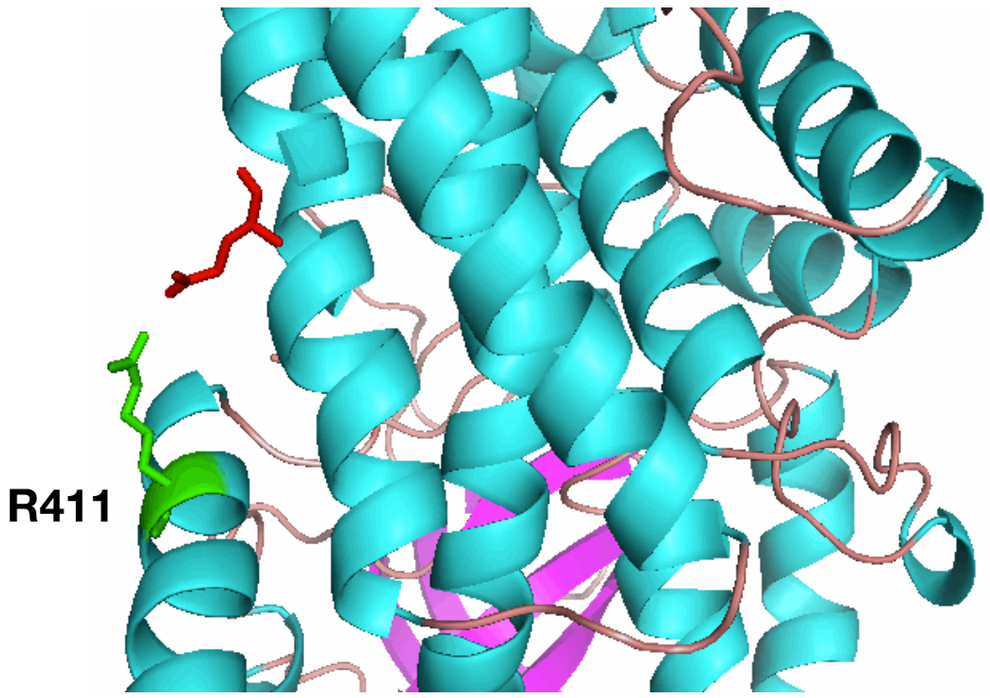


http://www.ncbi.nlm.nih.gov

Sequence, Structure & Function

Genomic variants in sequence motifs could affect protein function.
Mutation S362A of P53 affect the interaction with hydrolase USP7 and the deubiquitination of the protein.

Nonsynonymous variants responsible for protein structural changes and cause loss of stability of the folded protein.
Mutation R411L removes the salt bridge stabilizing the structure of the IVD dehydrogenase.



Machine learning

- Computational approach to build models based on the analysis of empirical data.

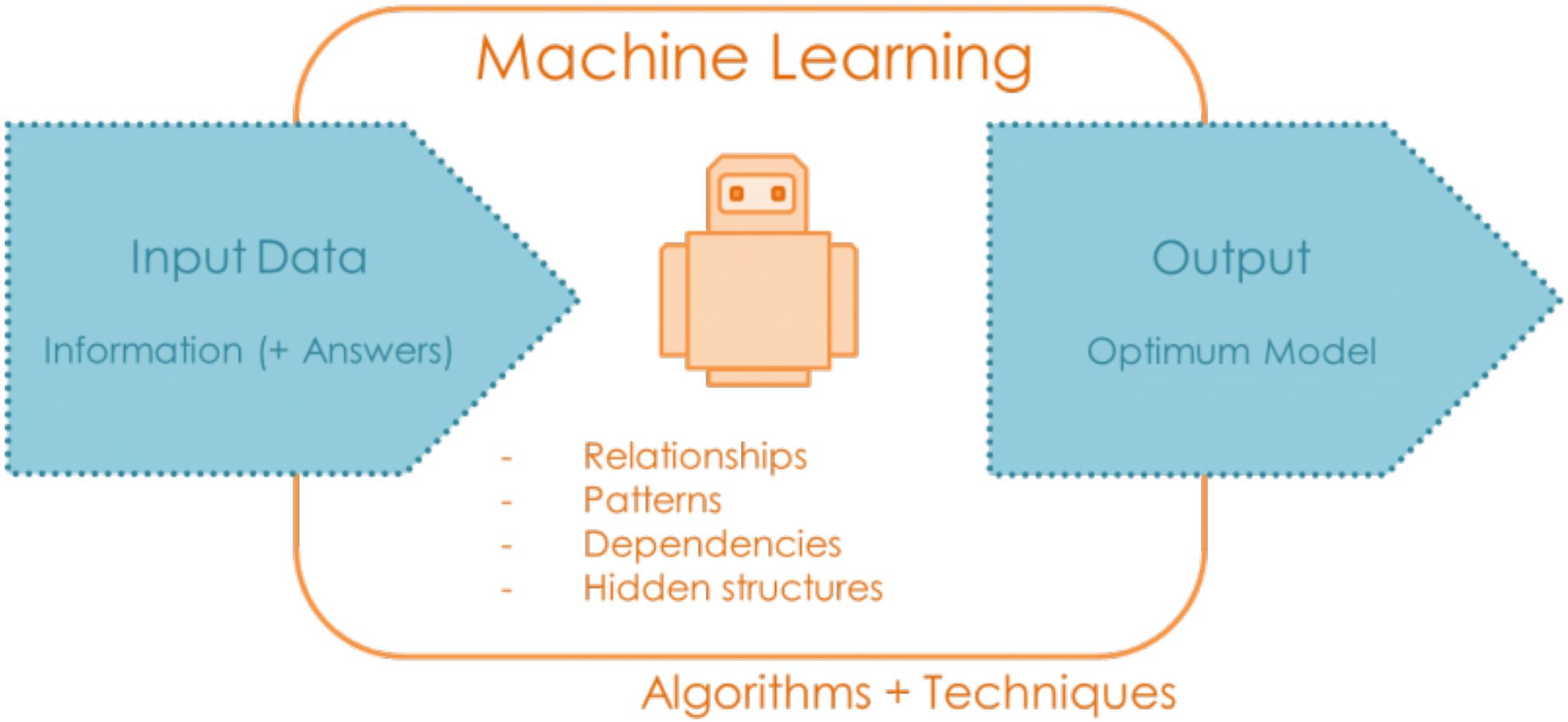
- Machine learning algorithms are suitable to address problems for which analytic solution does not exists and large amount of data are available.

- They are implemented selecting a representative set of data that are used in a training step and then validated on a test set with data *“not seen”* during the training.

- Most popular machine learning approaches are in computational biology are Neural Networks, Support Vector Machines and Random Forest.

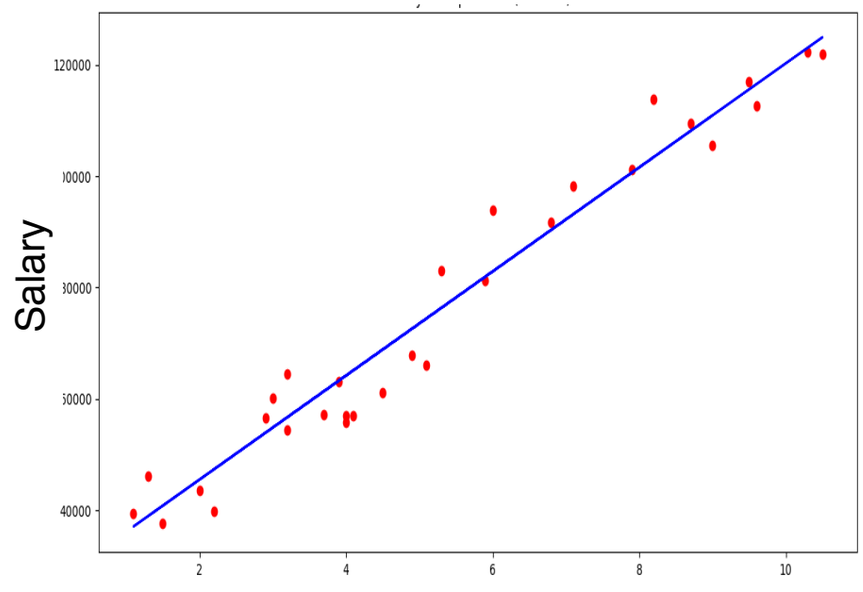
Input and Output

A machine learning algorithm takes in input a set of variables (features) and returns a numerical or discrete output

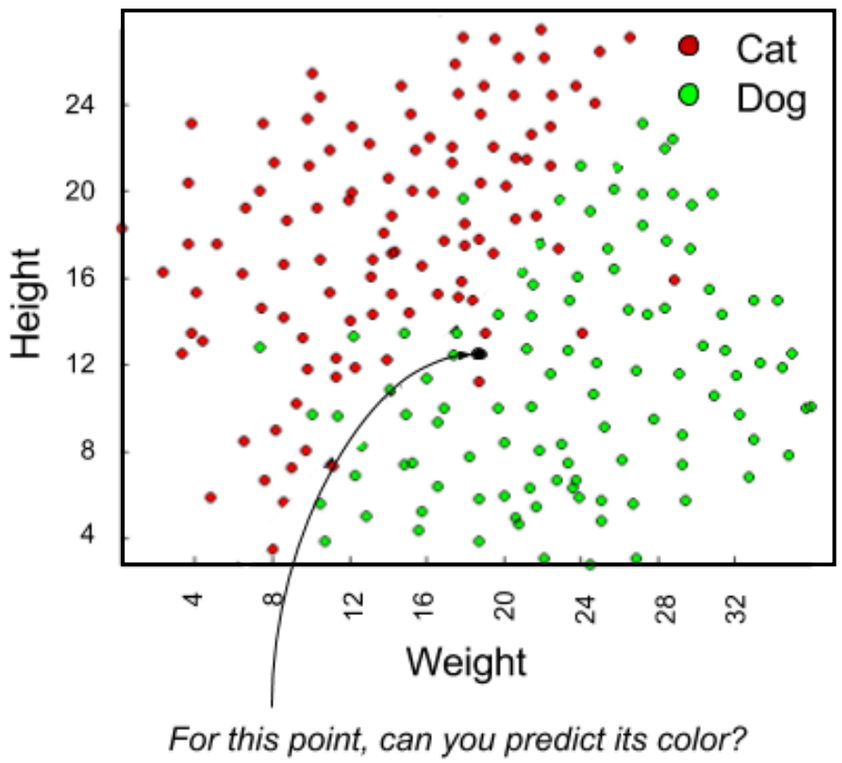


Types of Predictions

- Regression is used to predict continuous values.



- Classification is used to predict which class a data point is part of (discrete value).



Regression Evaluation

Compare predicted and real values using different correlation tests and the Root Mean Square Error

Pearson Correlation

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Root Mean Square Error

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}$$

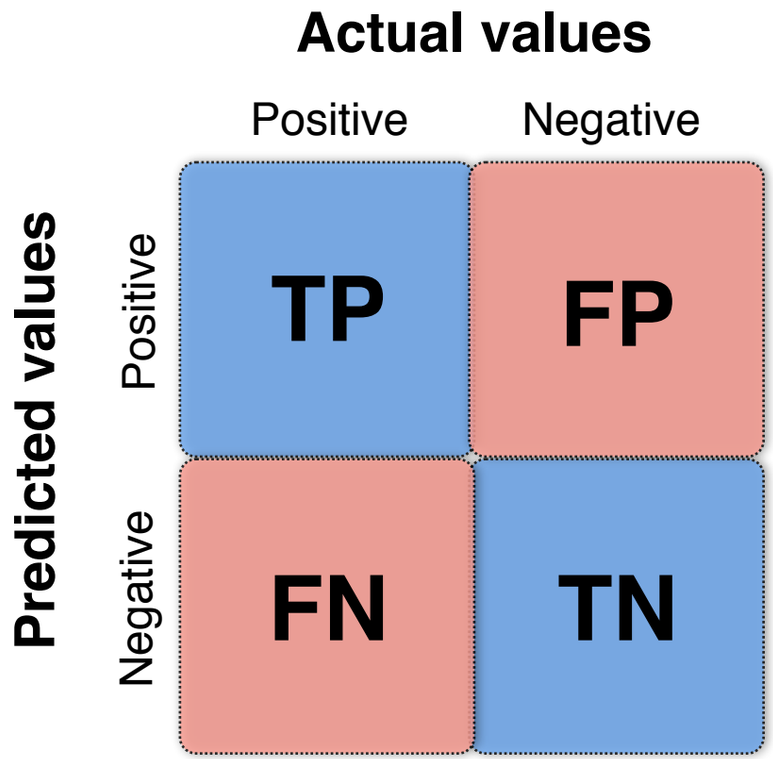
Classification Evaluation

Overall Accuracy

$$Q2 = \frac{TP + TN}{TP + FN + TN + FP}$$

Sensitivity

$$S = \frac{TP}{TP + FN}$$



Precision

$$P = \frac{TP}{TP + FP}$$

Matthews Correlation

$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

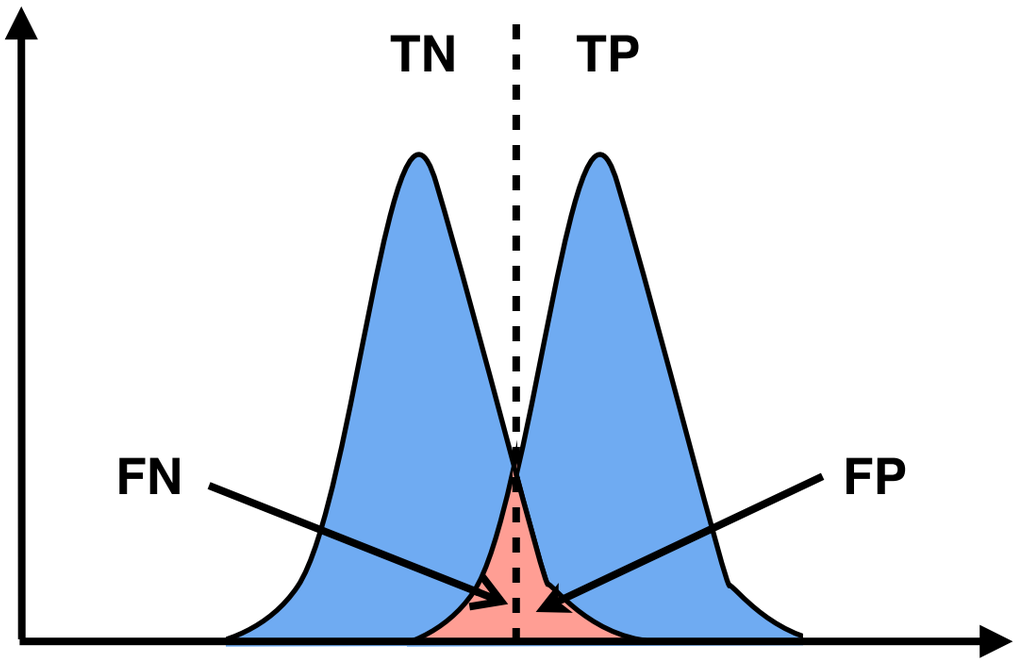
ROC Curve

True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

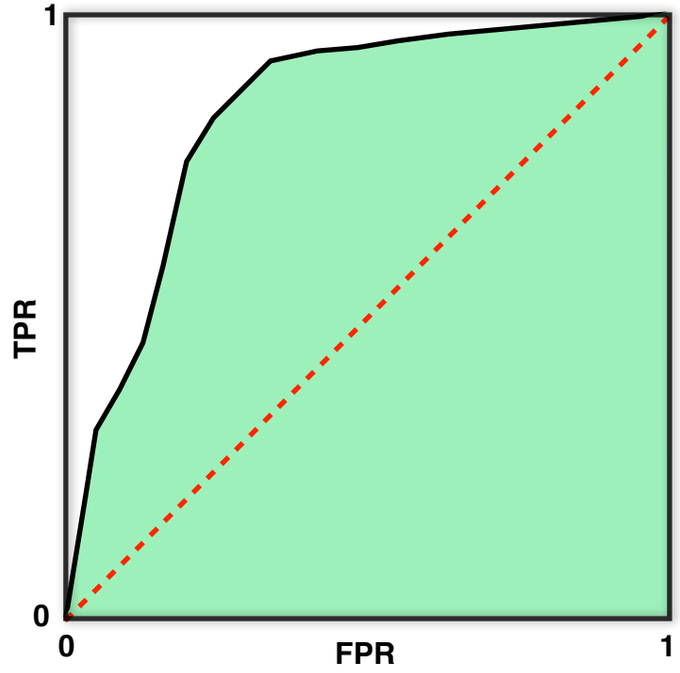
False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$



The Area Under the Receiver operating characteristic (ROC) Curve (AUC) is a prediction evaluation measure that is 0.5 for completely random predictors and close to 1.0 for highly accurate predictors.

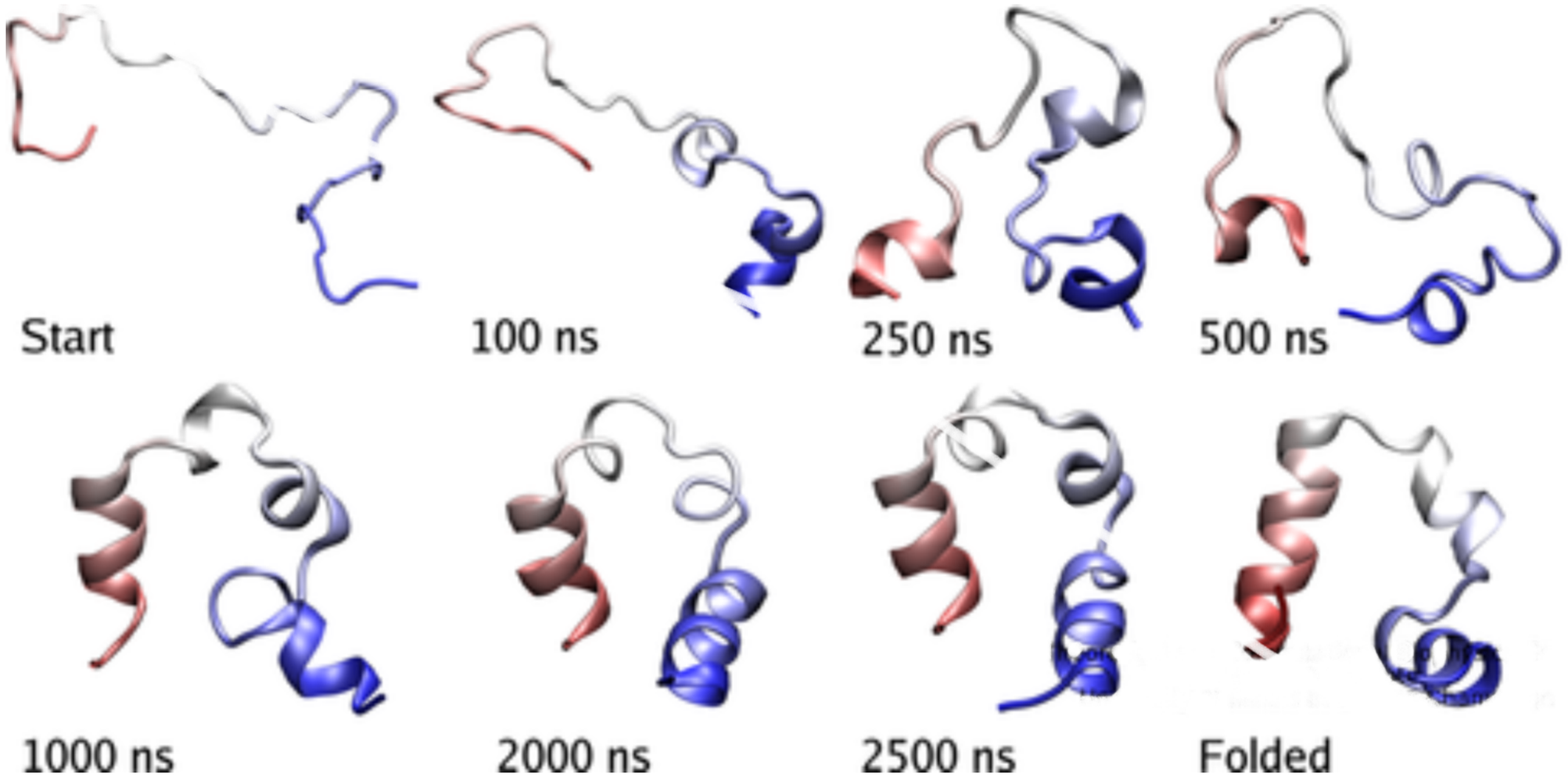
Baldi et al. (2000) Bioinformatics, 16:412-424



Mutation and Stability

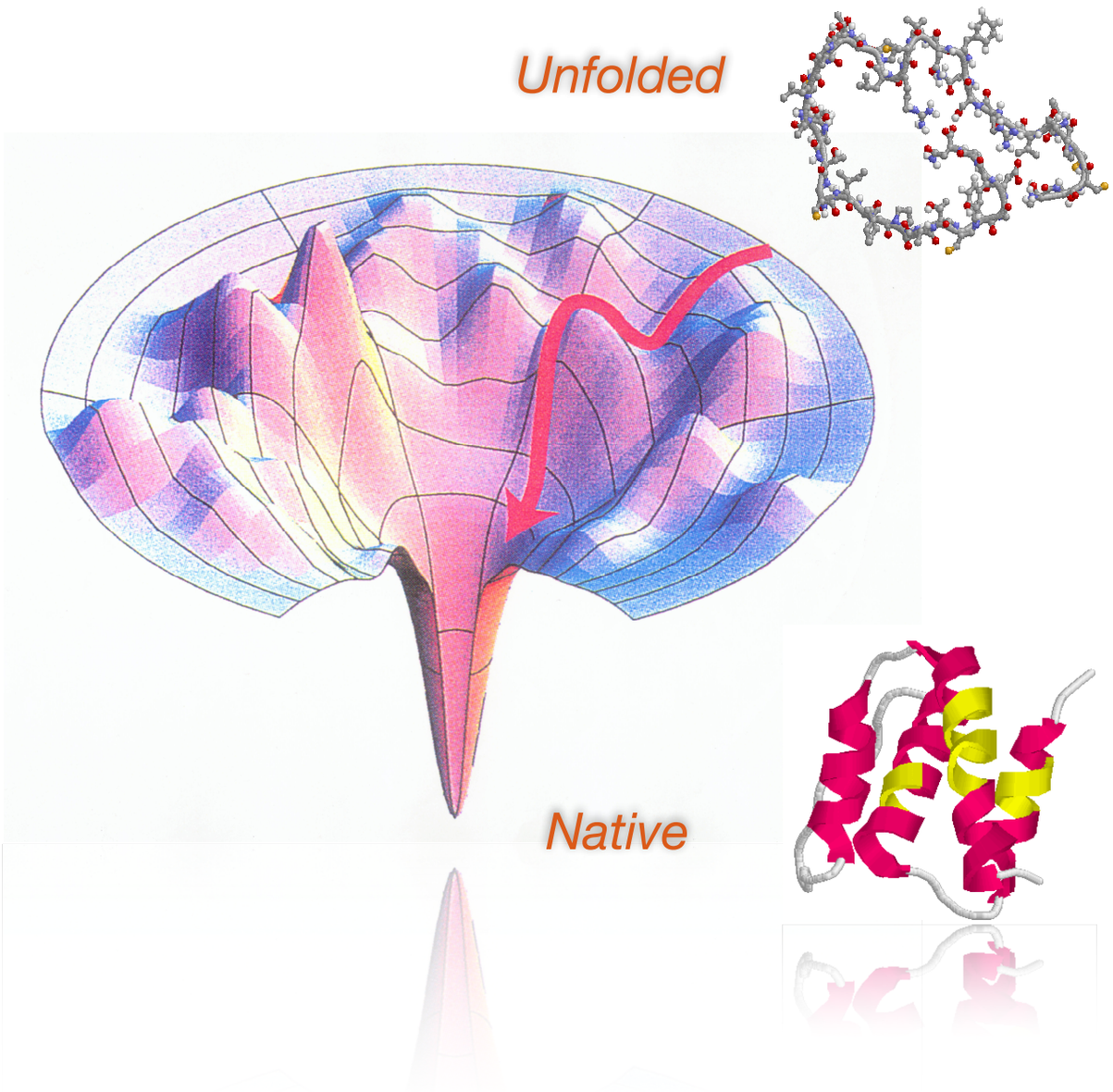
Protein folding

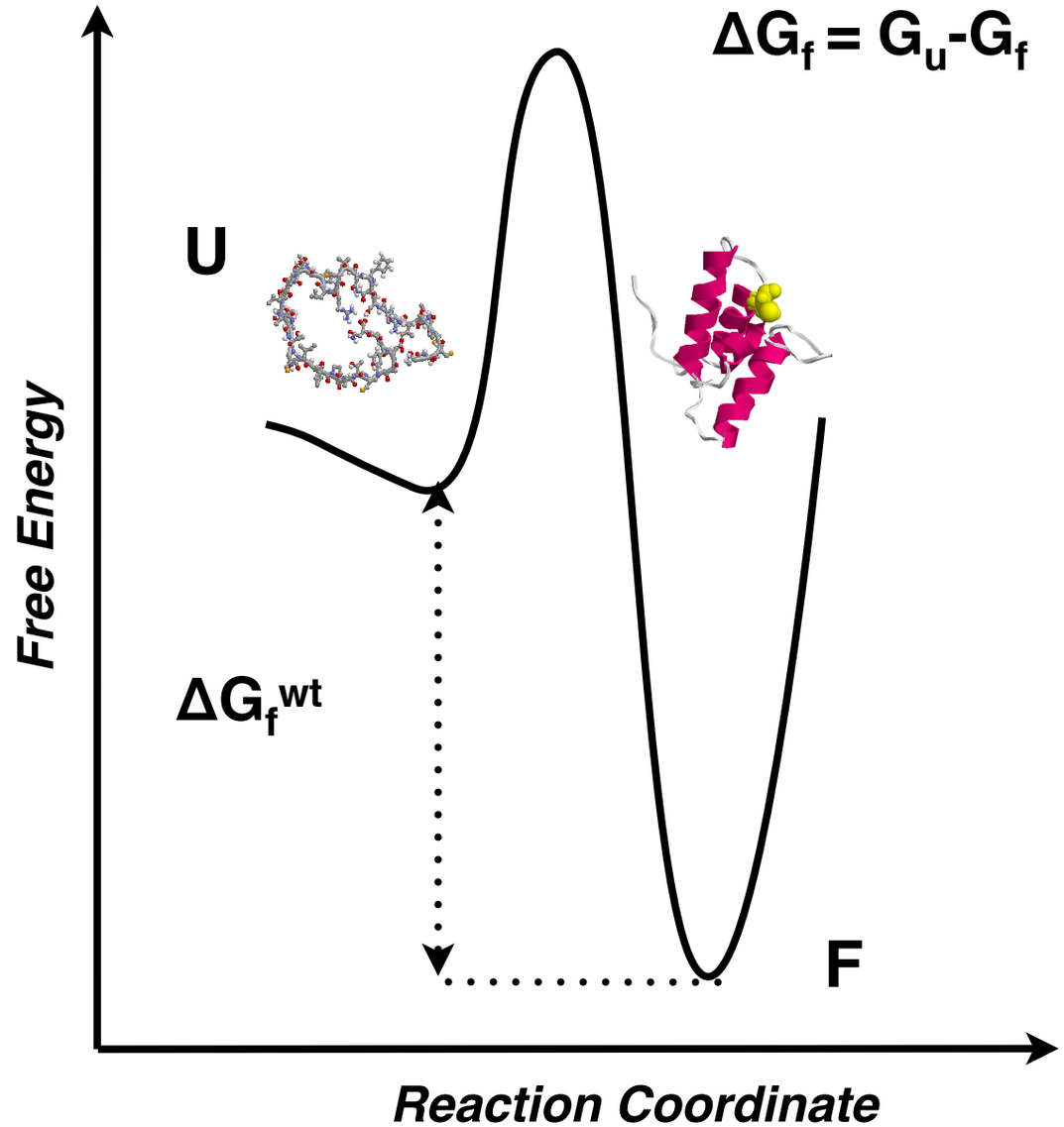
Protein folding is the process by which a protein assumes its native structure from the unfolded structure



Folding and stability

The folding free energy difference, ΔG_F, is typically small, of the order of -5 to -15 kcal/mol for a globular protein (compared to e.g. -30 to -100 kcal/mol for a covalent bond).





Folding and mutations

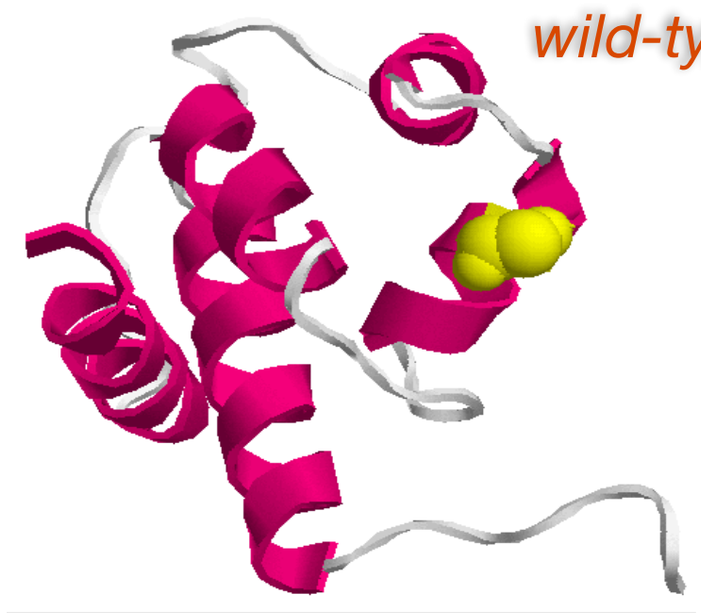
- Mutations of the protein sequence can affect the folding process changing the stability of the folded structure.

- Failure to folding process can produce inactive proteins with different properties even toxic. Protein misfolding is believed to be the main cause of neurodegenerative and other diseases.

- Web available databases are collecting large amount of thermodynamic data from mutagenesis experiments that can be used to develop methods for the prediction the protein stability change upon mutation.

Mutation and stability

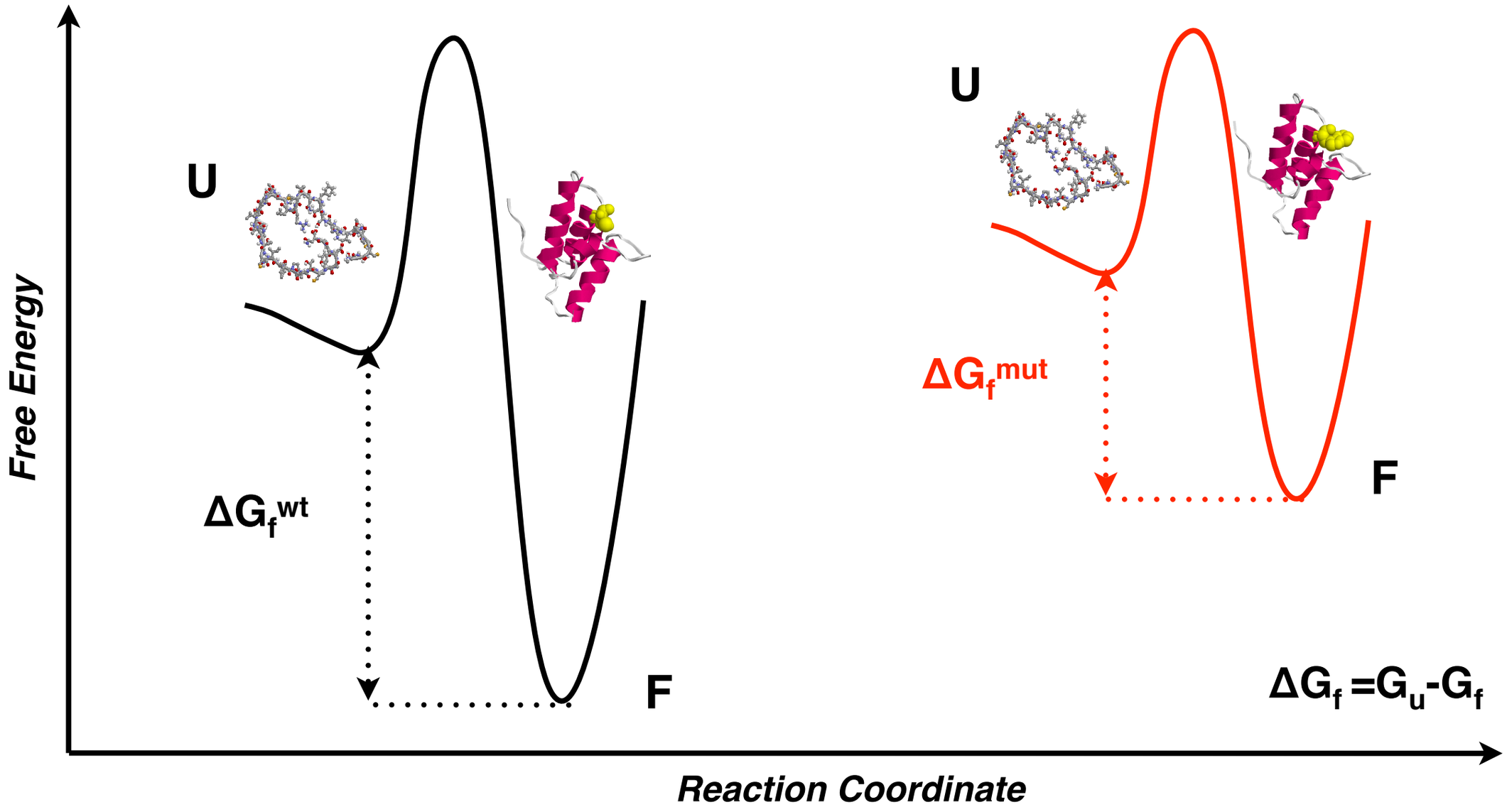
if a protein is mutated in a single site, what is the effect of the mutation on the stability of the protein?



Free energy change

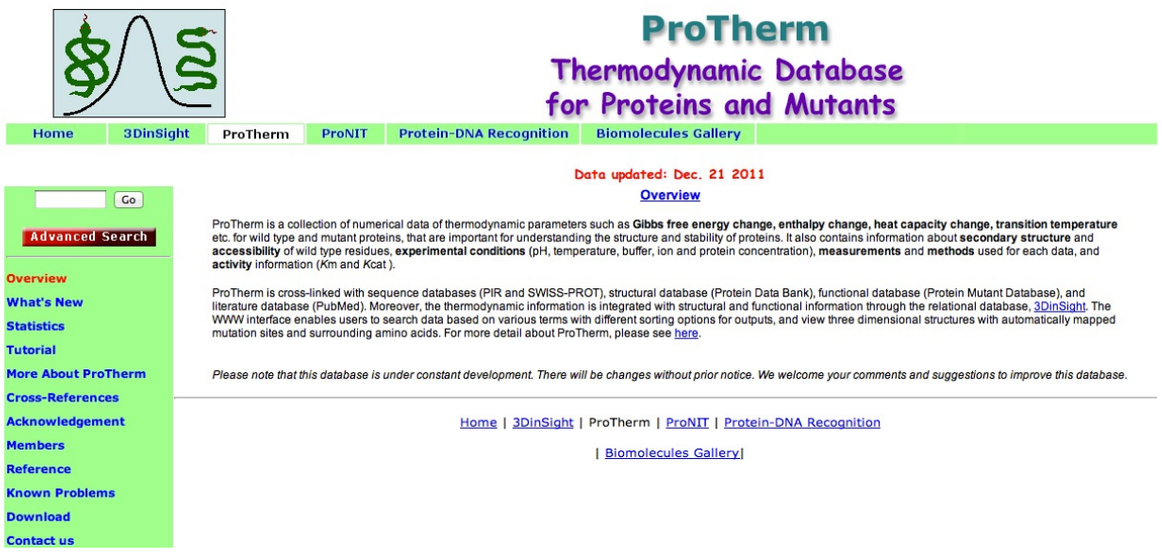
If we mutate one residue in the protein sequence,
is the protein stability increased or decreased?

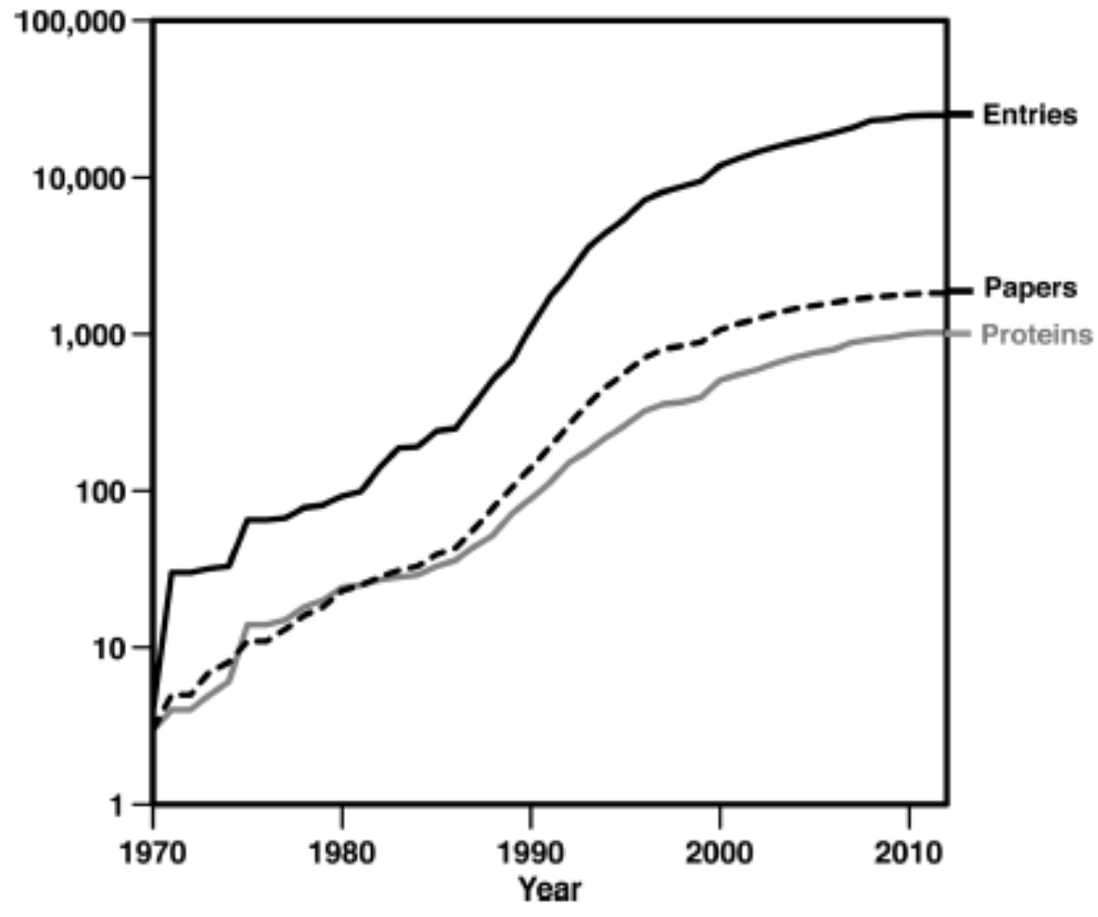
$$\Delta\Delta G_f = \Delta G_f^{mut} - \Delta G_f^{wt}$$



ProTherm database

ProTherm is a collection of numerical data of thermodynamic parameters including Gibbs free energy change, enthalpy change, heat capacity change, transition temperature etc. for wild type and mutant proteins, that are important for understanding the structure and stability of proteins.

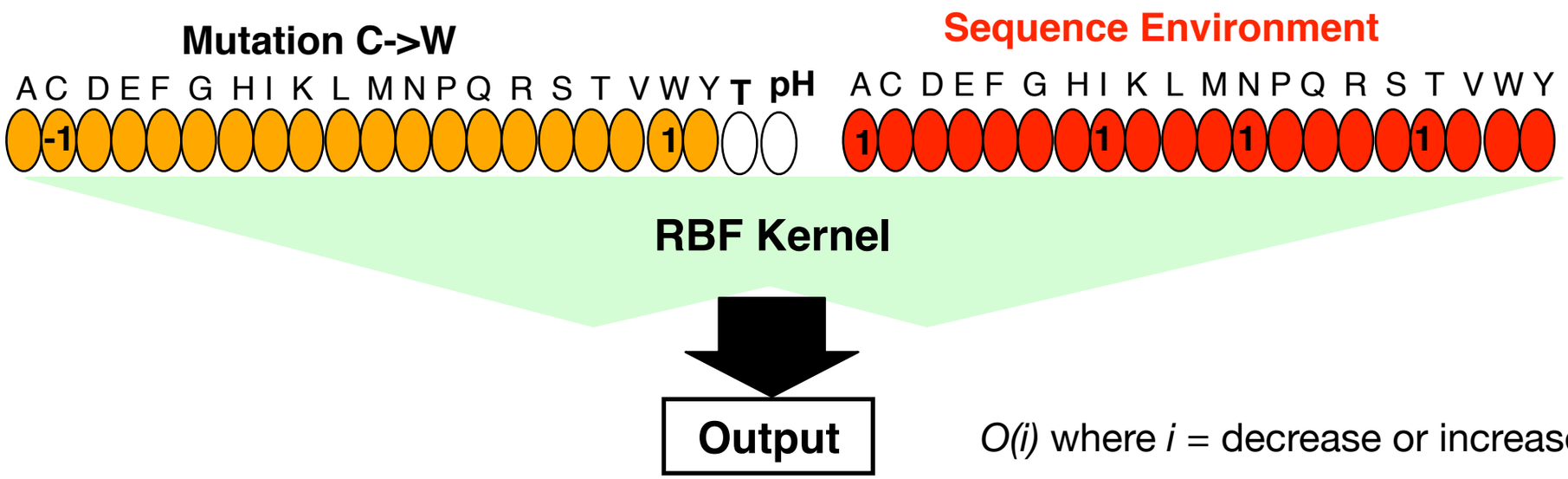




Total number of entries	**25820**
Number of unique proteins	740
Total number of all proteins	1045
Number of Proteins with mutants	311
Number of Single Mutations	**12561**
Number of Double Mutations	1744
Number of Multiple Mutations	1132
Number of Wild Type	10383

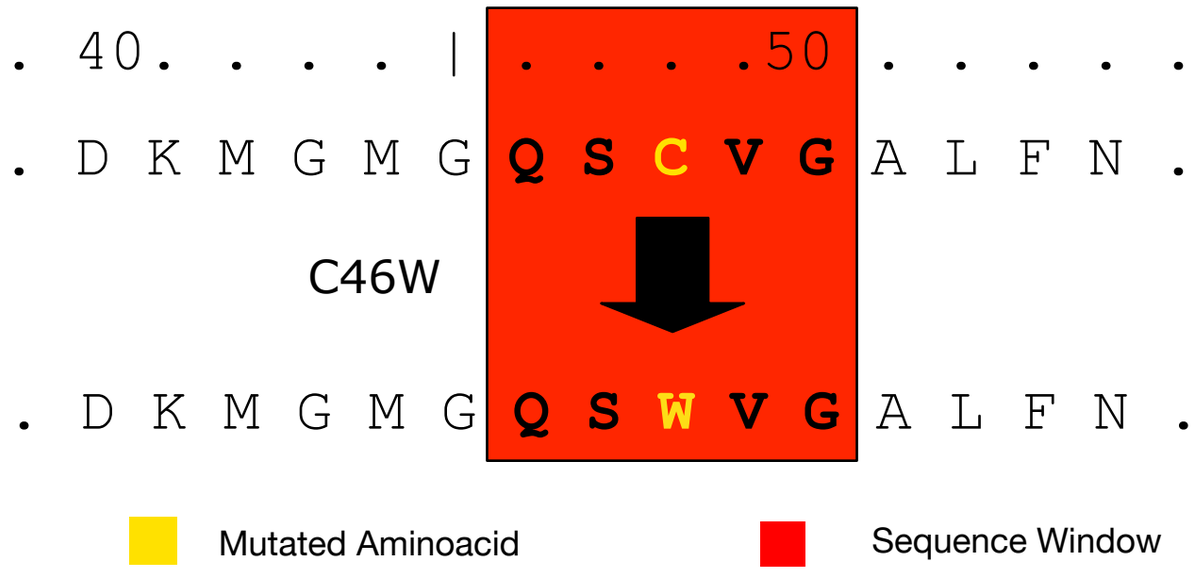
Feb 2013

Sequence-based predictor



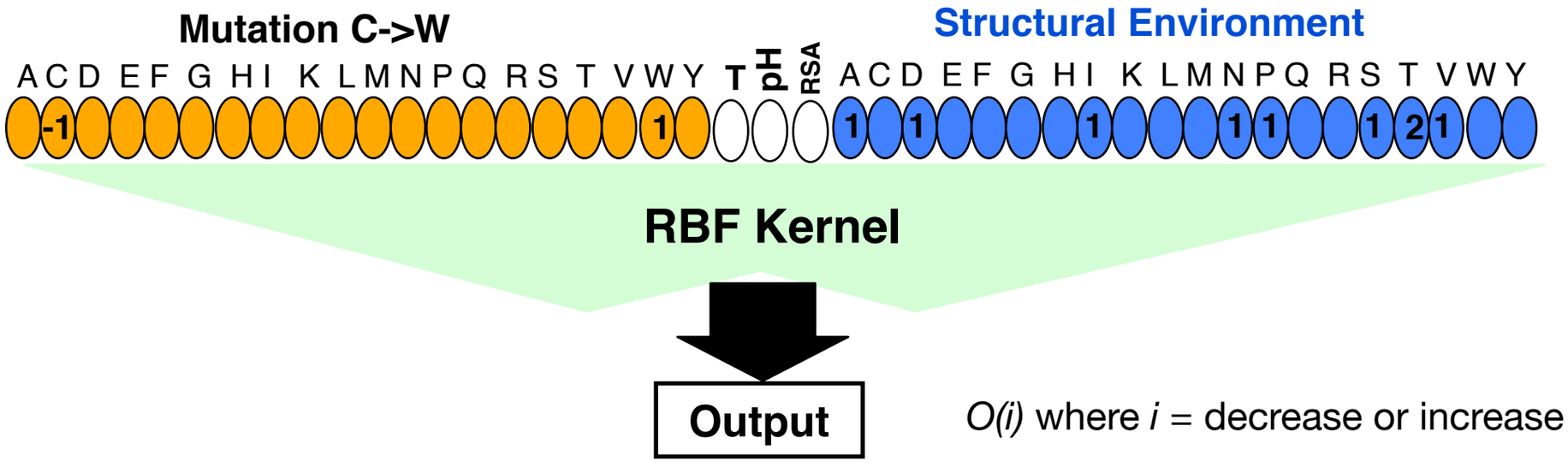
O(i) where *i* = decrease or increase stability

SVM-SEQUENCE: 20 element vector that describes the amino acid mutation,
2 element pH and T (experimental conditions)
20 more input features (40 in total) encoding the sequence residue environment



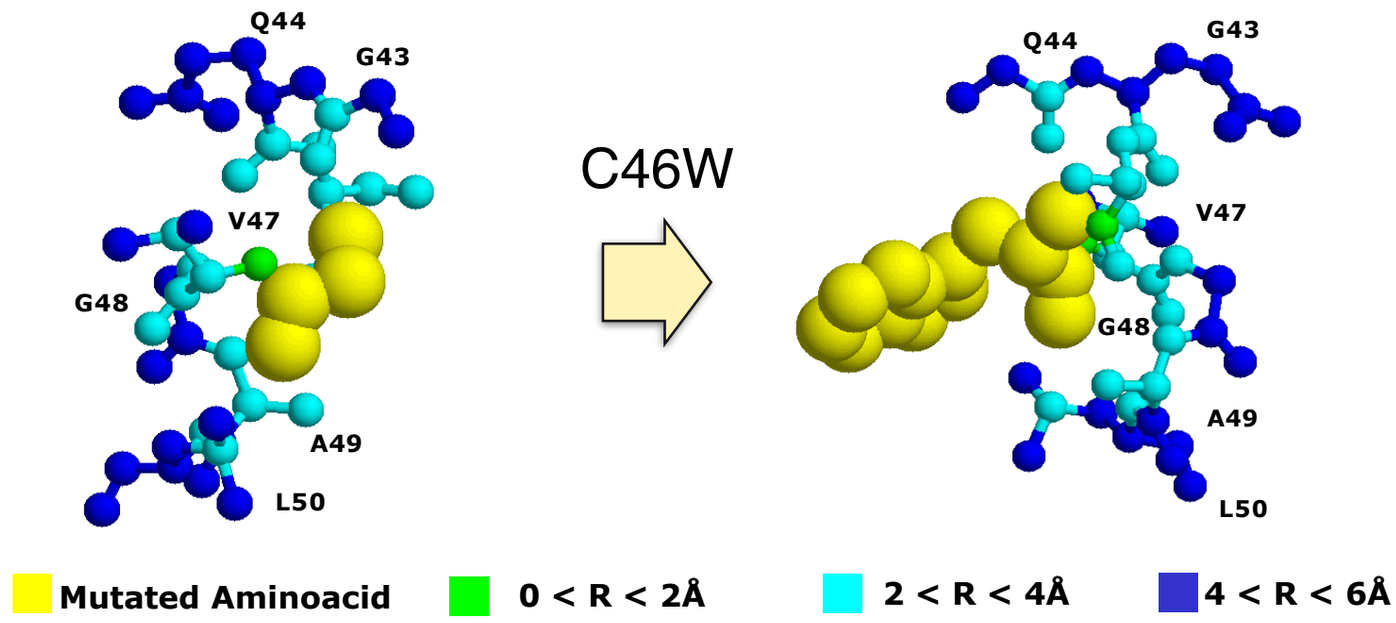
Capriotti et al. (2005) Bioinformatics, 21: ii54-ii58.

Structure-based predictor



O(i) where *i* = decrease or increase stability

SVM-STRUCTURE: 20 element vector that describes the amino acid mutation,
3 element pH, T and relative solvent accessible area
20 more input features (43 in total) encoding the structure residue environment

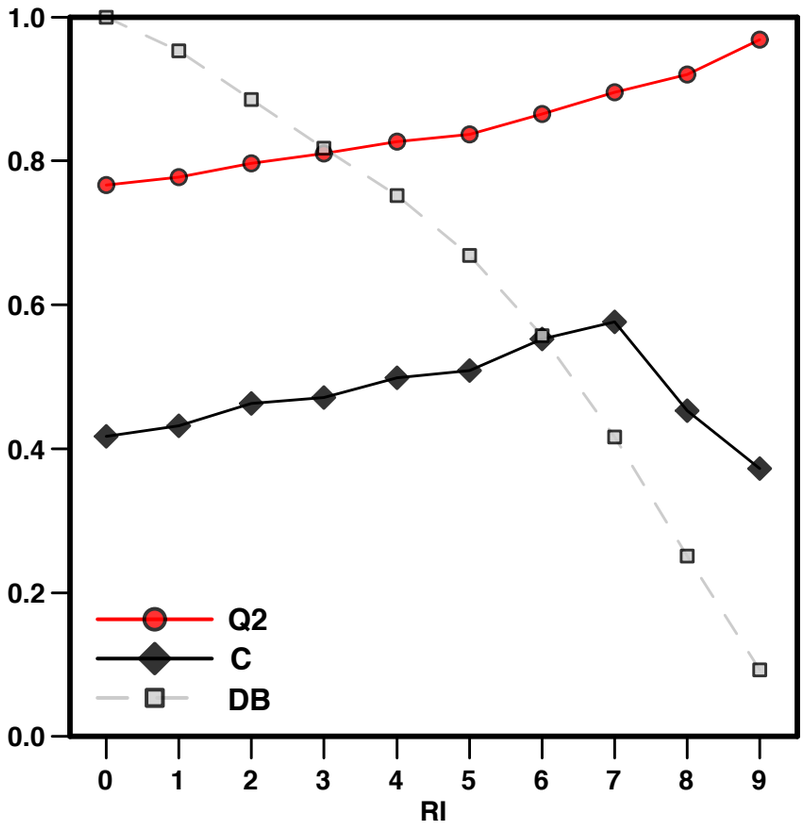


Classification results

	Q2	P[-]	S[-]	P[+]	S[+]	C
SVM-Sequence	0.77	0.79	0.91	0.69	0.46	0.42
SVM-Structure	0.80	0.83	0.91	0.73	0.56	0.51

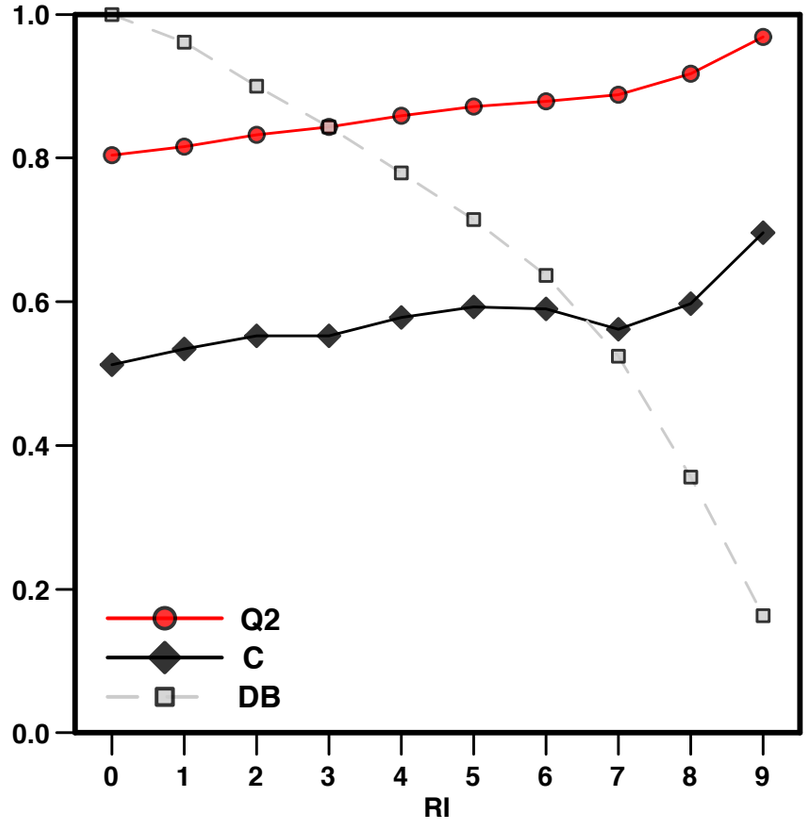
+ Increase stability – Decrease stability

Sequence-based predictor



DBSEQ= 2087 mutations

Structure-based predictor

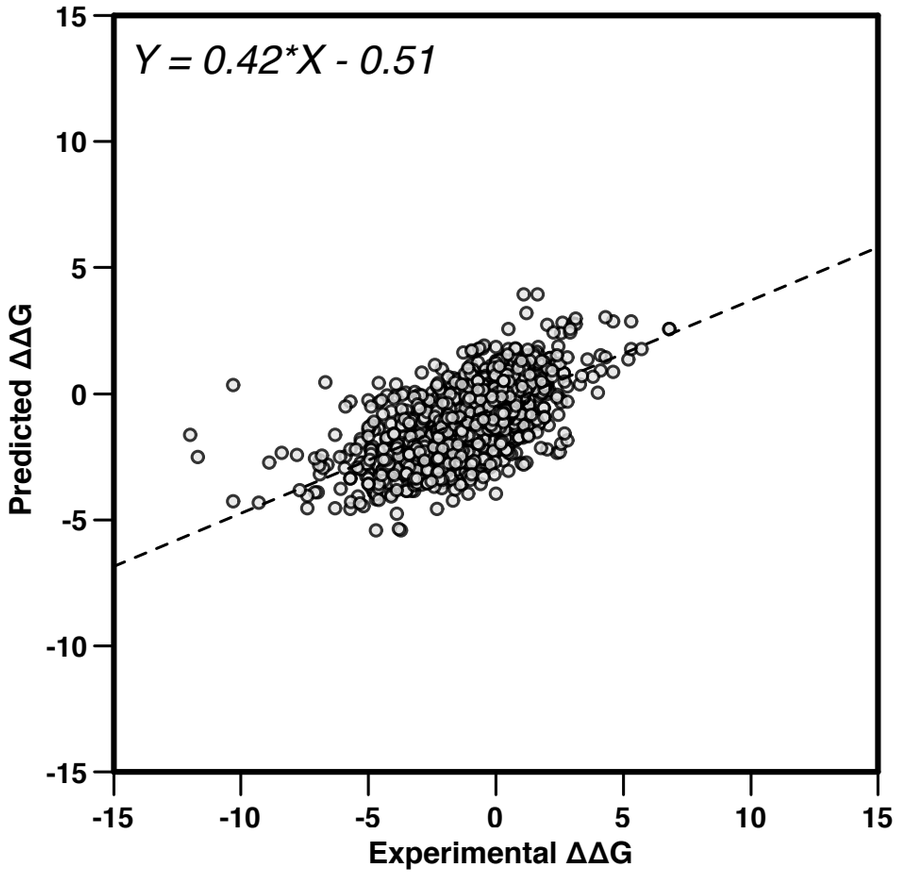


DB3D= 1948 mutations

Q2: Overall Accuracy **C:** Mean Correlation Coefficient **DB:** Fraction of database that are predicted with a reliability the given threshold

Regression results

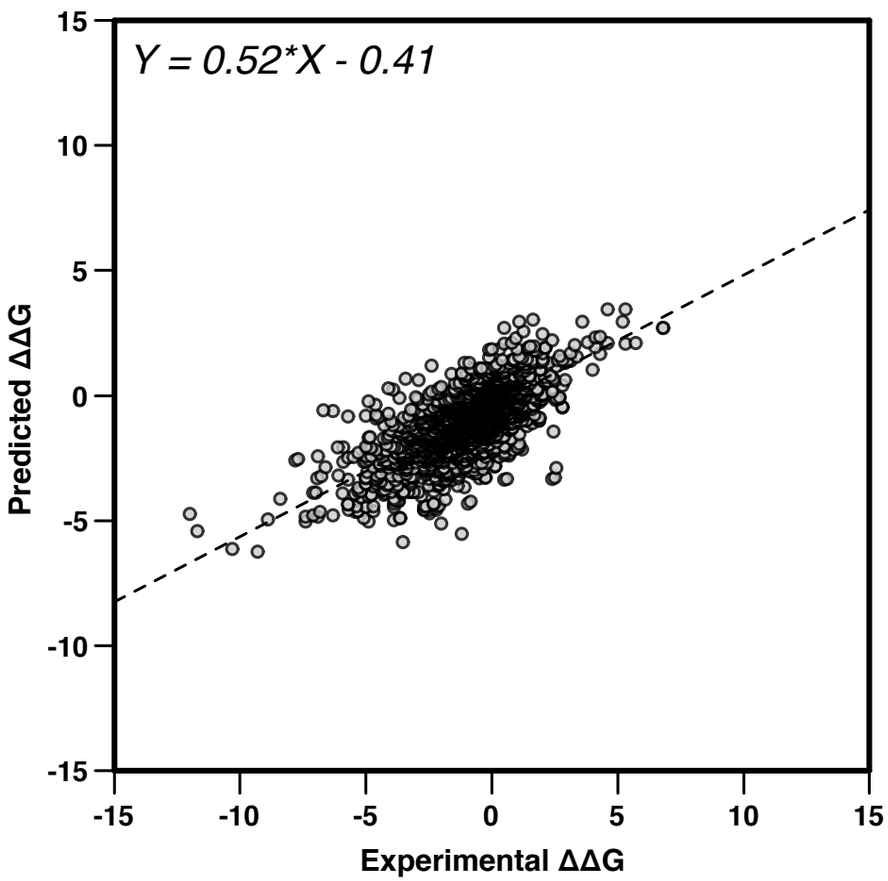
Sequence-based predictor



DBSEQ= 2087 mutations

C= 0.62 (RMSE= 1.45 kcal/mole)

Structure-based predictor



DB3D= 1948 mutations

C= 0.71 (RMSE= 1.30 kcal/mole)

http://folding.biofold.org/i-mutant

Capriotti et al. (2005) Nucleic Acids Research 33, W306-W310.

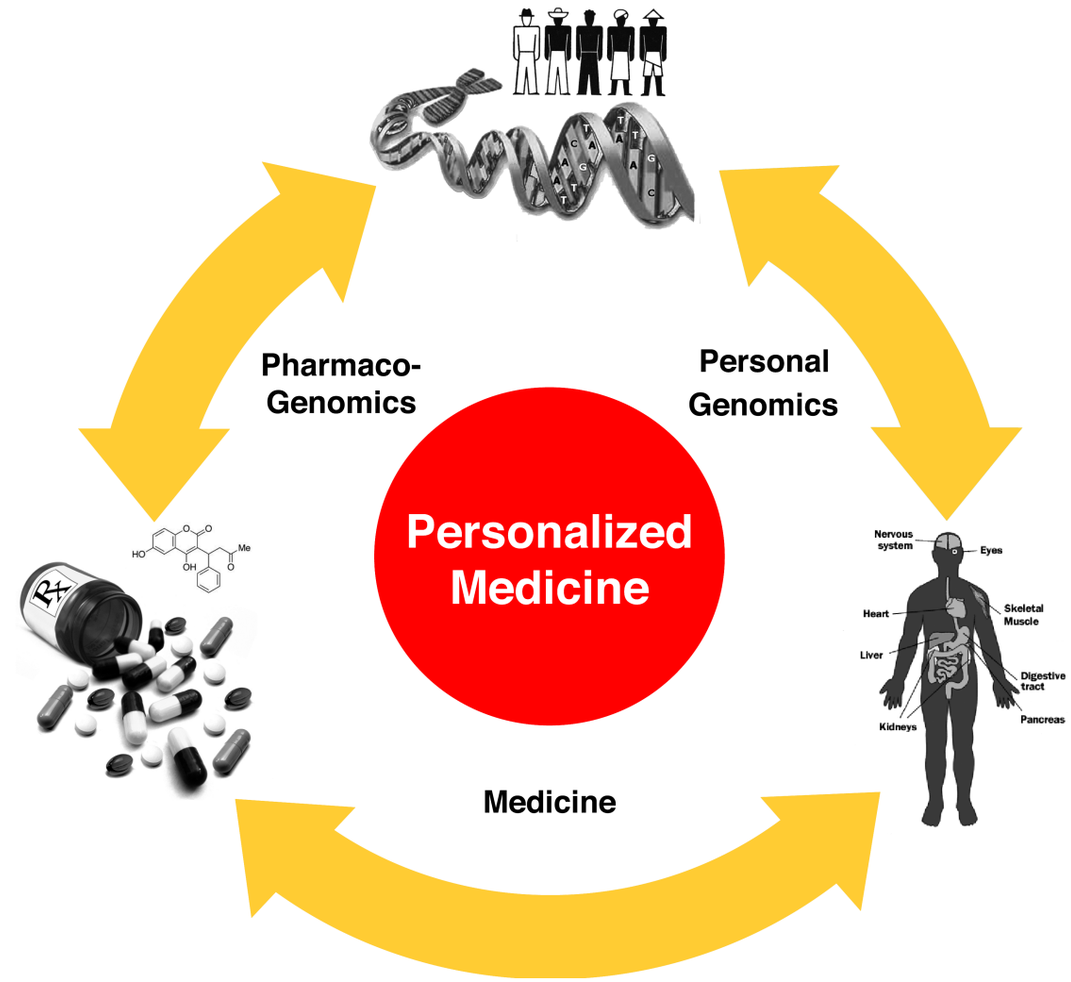
Mutation and Disease

Personalized medicine

Currently direct to consumers company are performing genotype test on markers associated to genetic traits, and soon full genome sequencing will cost ~$1000.

The future bioinformatics challenges for personalized medicine will be:

1. Processing Large-Scale Robust Genomic Data
2. Interpretation of the Functional Effect and the Impact of Genomic Variation
3. Integrating Systems and Data to Capture Complexity
4. Making it all clinically relevant



Fernald GH, et al (2011). Bioinformatics. 27: 1741-1748.

1000 Genomes

The 1000 Genomes Project aims to create the largest public catalogue of human variations and genotype data. Last versione released the genotype of ~2,500 individuals.

Table 1 | Variants discovered by project, type, population and novelty

a Summary of project data including combined exon populations

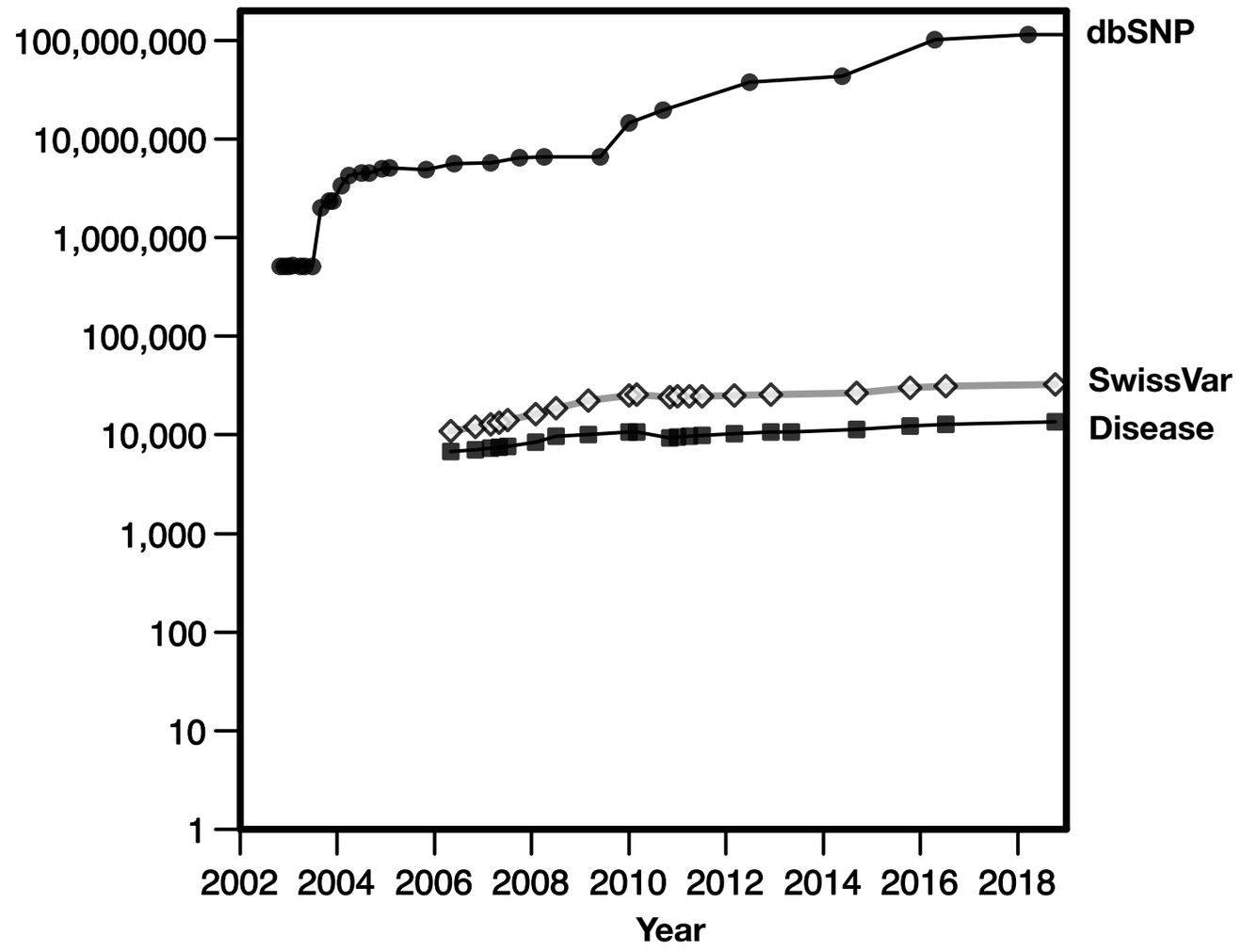
	Low coverage				Trios				
Statistic	CEU	YRI	CHB+JPT	Total	CEU	YRI	Total	Exon (total)	Union across projects
Samples	60	59	60	179	3	3	6	697	742
Total raw bases (Gb)	1,402	874	596	2,872	560	615	1,175	845	4,892
Total mapped bases (Gb)	817	596	468	1,881	369	342	711	56	2,648
Mean mapped depth (×)	4.62	3.42	2.65	3.56	43.14	40.05	41.60	55.92	NA
Bases accessed (% of genome)	2.43 Gb (86%)	2.39 Gb (85%)	2.41 Gb (85%)	2.42 Gb (86.0%)	2.26 Gb (79%)	2.21 Gb (78%)	2.24 Gb (79%)	1.4 Mb	NA
No. of SNPs (% novel)	7,943,827 (33%)	10,938,130 (47%)	6,273,441 (28%)	14,894,361 (54%)	3,646,764 (11%)	4,502,439 (23%)	5,907,699 (24%)	12,758 (70%)	15,275,256 (55%)
Mean variant SNP sites per individual	2,918,623	3,335,795	2,810,573	3,019,909	2,741,276	3,261,036	3,001,156	763	NA
No. of indels (% novel)	728,075 (39%)	941,567 (52%)	666,639 (39%)	1,330,158 (57%)	411,611 (25%)	502,462 (37%)	682,148 (38%)	96 (74%)	1,480,877 (57%)
Mean variant indel sites per individual	354,767	383,200	347,400	361,669	322,078	382,869	352,474	3	NA
No. of deletions (% novel)	ND	ND	ND	15,893 (60%)	6,593 (41%)	8,129 (50%)	11,248 (51%)	ND	22,025 (61%)
No. of genotyped deletions (% novel)	ND	ND	ND	10,742 (57%)	ND	ND	6,317 (48%)	ND	13,826 (58%)
No. of duplications (% novel)	259 (90%)	320 (90%)	280 (91%)	407 (89%)	187 (93%)	192 (91%)	256 (92%)	ND	501 (89%)
No. of mobile element insertions (% novel)	3,202 (79%)	3,105 (84%)	1,952 (76%)	4,775 (86%)	1,397 (68%)	1,846 (78%)	2,531 (78%)	ND	5,370 (87%)
No. of novel sequence insertions (% novel)	ND	ND	ND	ND	111 (96%)	66 (86%)	174 (93%)	ND	174 (93%)

1000 Genomes Project Consortium (2010). Nature. 467: 1061-1073.

SNVs and Disease

Single Nucleotide Variants (SNVs) are the most common type of genetic variations in human accounting for more than 90% of sequence differences (1000 Genome Project Consortium, 2012).

SNVs can also be responsible of genetic diseases (Ng and Henikoff, 2002; Bell, 2004).

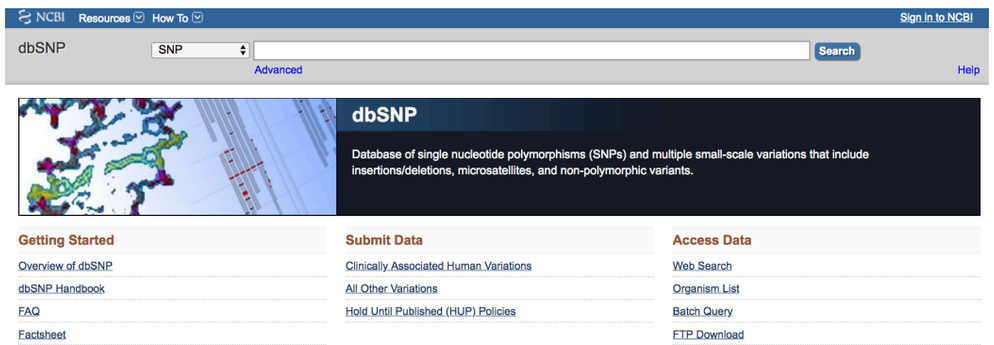


nonsynonymous SNVs

- disease-related SNVs → the mutations are related to a Mendelian pathologies
- neutral SNVs → the mutations do not compromise the organism's health

SNVs and SAVs databases

dbSNP (Mar 2018) @ NCBI



http://www.ncbi.nlm.nih.gov/snp

Single Nucleotide Variants

Homo sapiens	**113,862,023**
Gallus gallus	15,104,956
Zea mays	14,672,946

SwissVar (Oct 2018) @ ExPASy

SIB

swissvar

http://www.expasy.ch/swissvar/

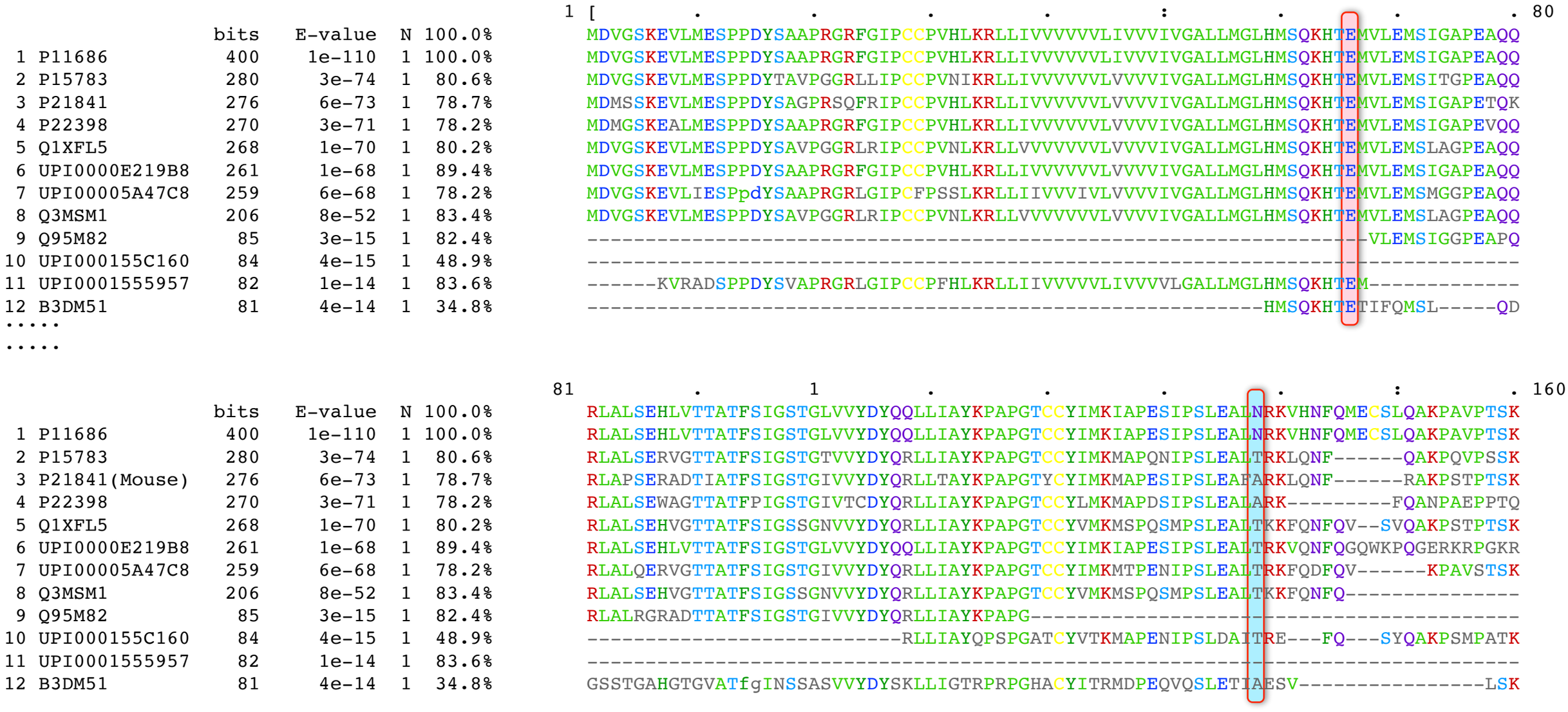
Single Amino acid Variants

Homo sapiens	76,608
Disease	*29,529*
Polymorphisms	*39,779*

Oct 2018

Conserved or not?

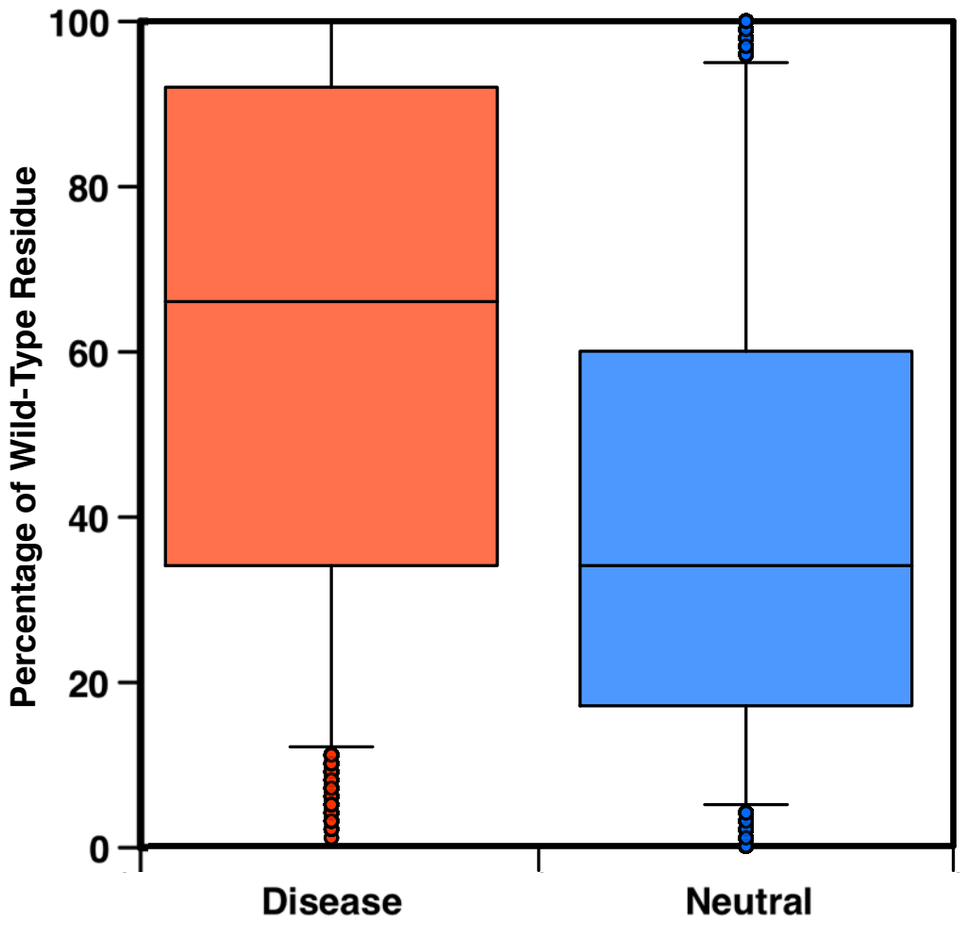
In positions 66 the Glutamic acid is highly conserved Asparagine in position 138 is mutated Threonine or Alanine



Sequence profile

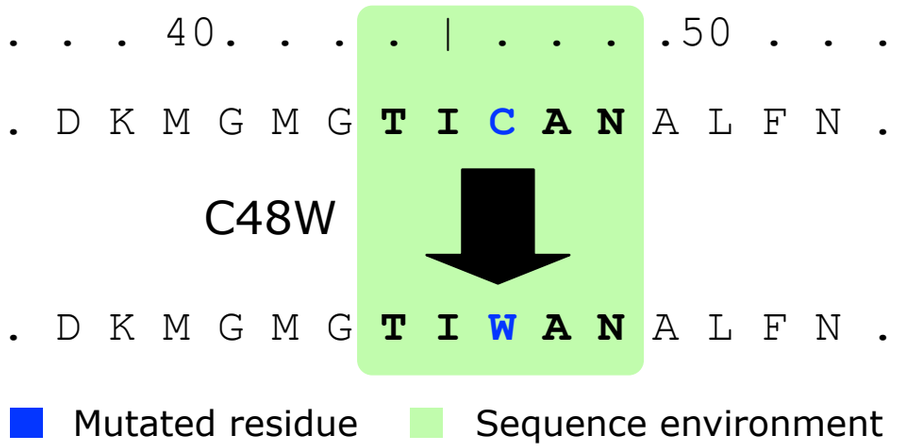
The protein sequence profile is calculated running BLAST on the UniRef90 dataset and selecting only the hits with e-value < 10^{-9}.

The frequency distributions of the wild-type residues for disease-related and neutral variants are significantly different (KS p-value=0).

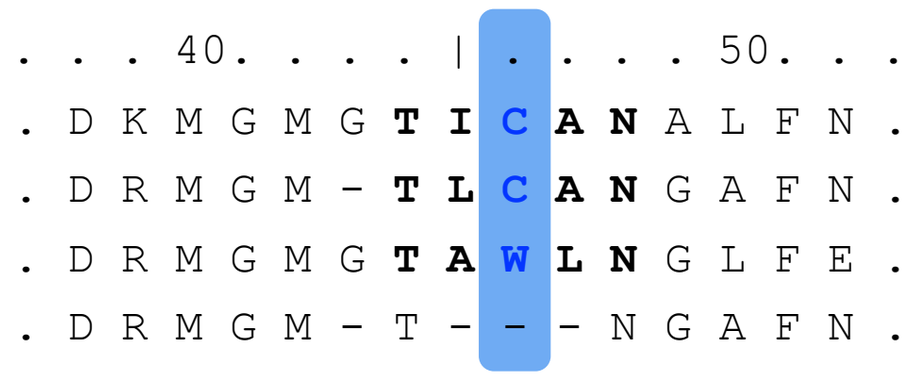


Capriotti et al (2012). Briefings in Bioinformatics. 13; 495-512.

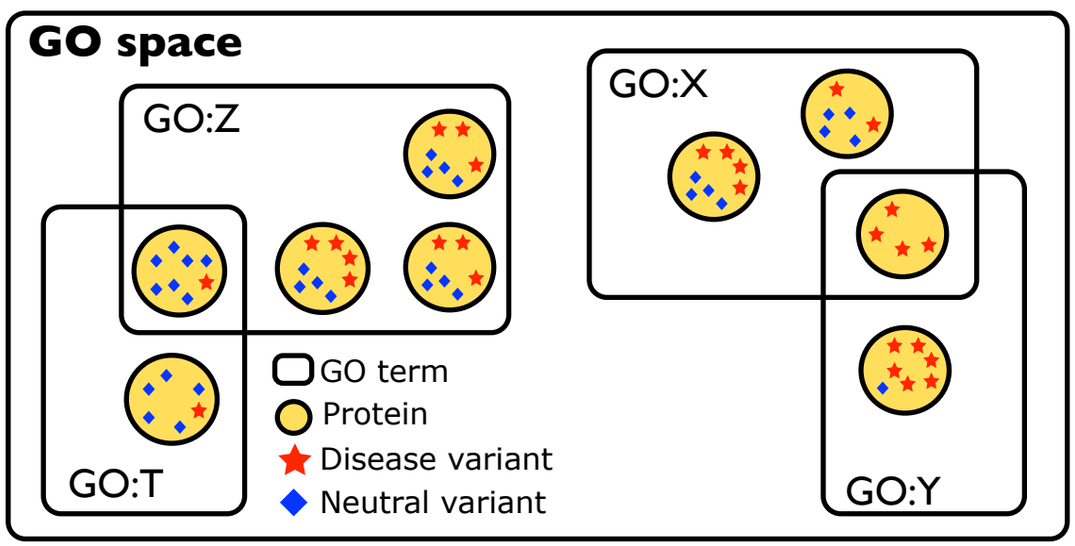
SNPs&GO input features



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



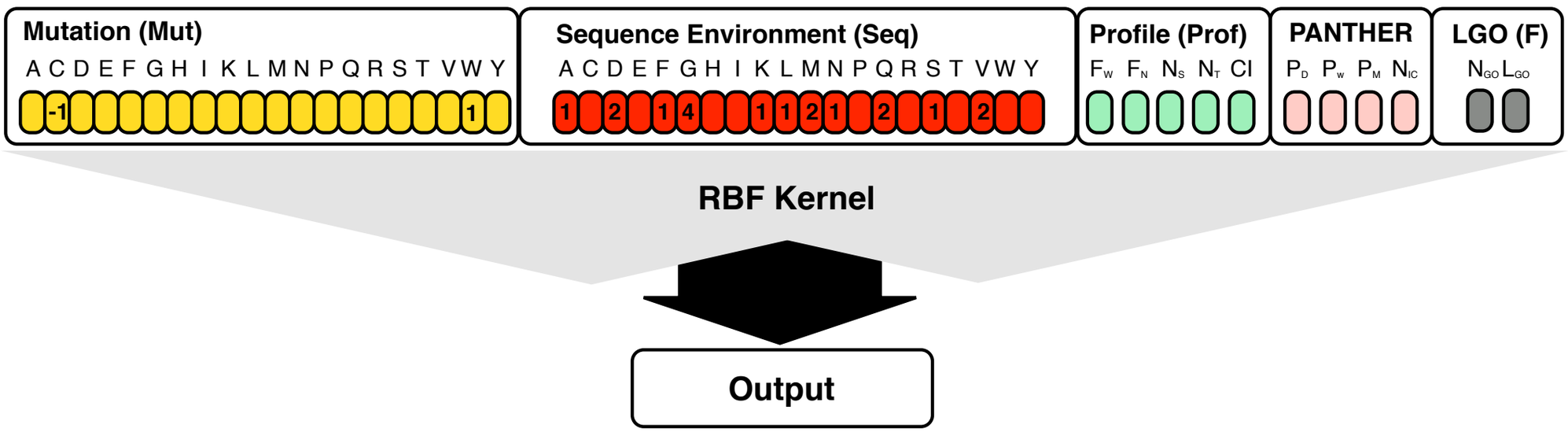
Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features



The GO information are encoded in a 2 elements vector corresponding to the number unique of GO terms associated to the protein sequences and the sum of the logarithm of the total number of disease-related and neutral variants for each GO term.

SNPs&GO performance

SNPs&GO results in better performance with respect to previously developed methods.



Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
PolyPhen	0,71	0,76	0,75	0,63	0,64	0,39	58
SIFT	0,76	0,75	0,76	0,77	0,75	0,52	93
PANTHER	0,74	0,77	0,73	0,71	0,76	0,48	76
SNPs&GO	0.82	0.83	0.78	0.80	0.85	0.63	100

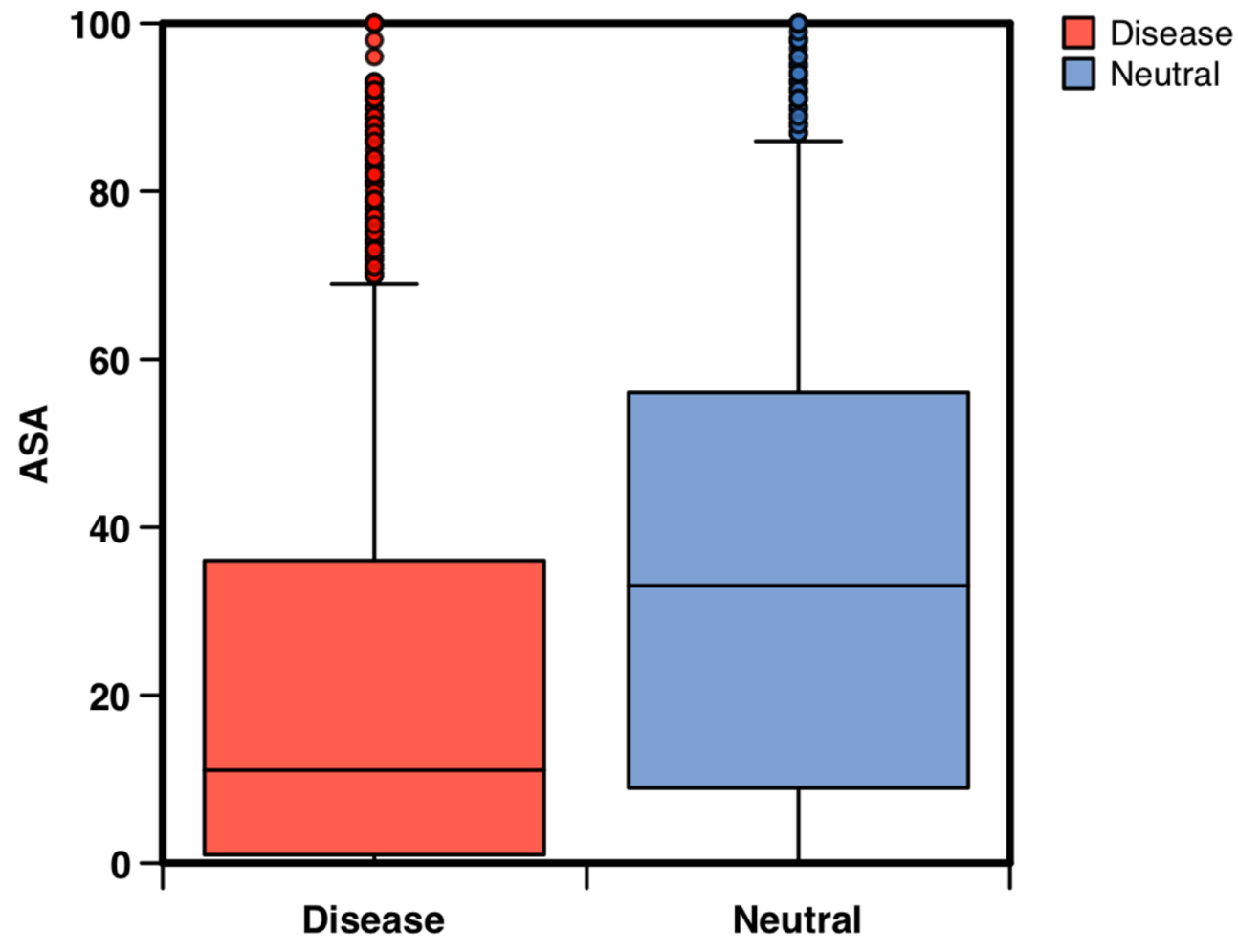
D = Disease related N = Neutral

DB= 33672 nsSNVs

Calabrese et al. (2009) Human Mutation 30, 1237-1244.

Structure environment

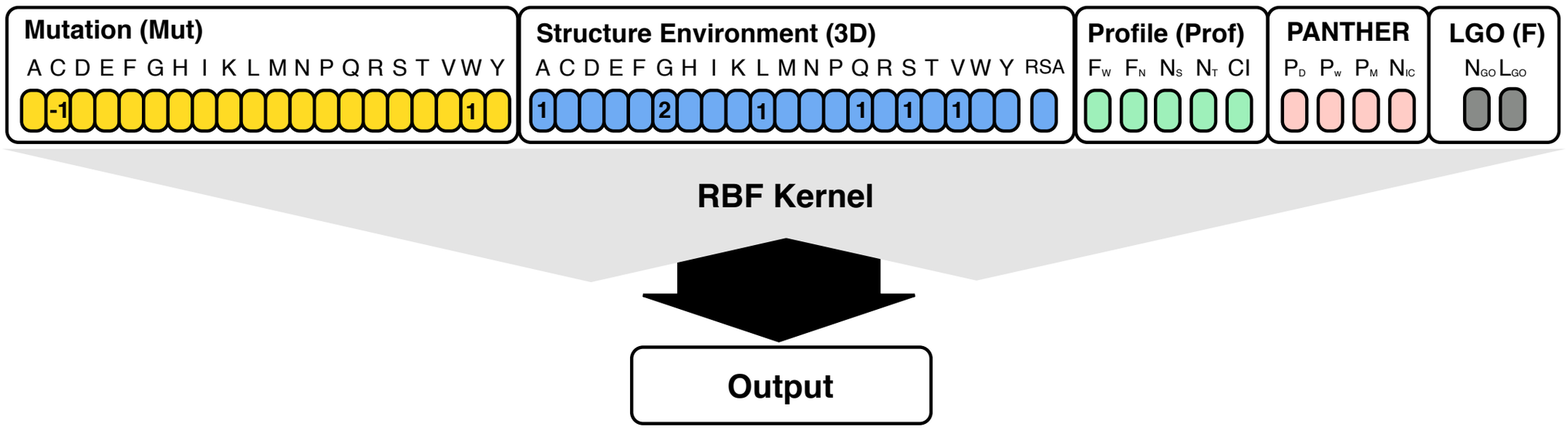
There is a significant difference (KS p-value = 2.8×10^{-71}) between the distributions of the relative Accessible Solvent Area for disease-related and neutral variants. Their mean values are respectively 20.6 and 35.7.

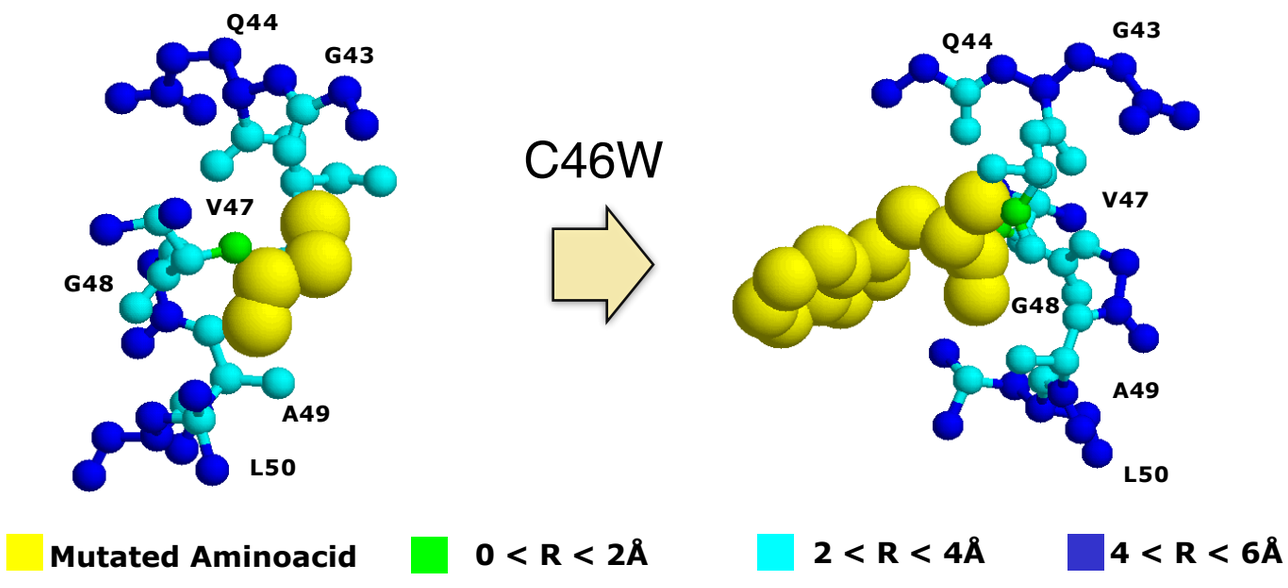


Capriotti and Altman. (2011) BMC Bioinformatics.12 (Suppl 4); S3.

The structure-based method

The method takes in to input 4 types of information encoded in a 48 elements vector.
The input features are: mutation data; structure environment, sequence profile and functional score based on GO terms.

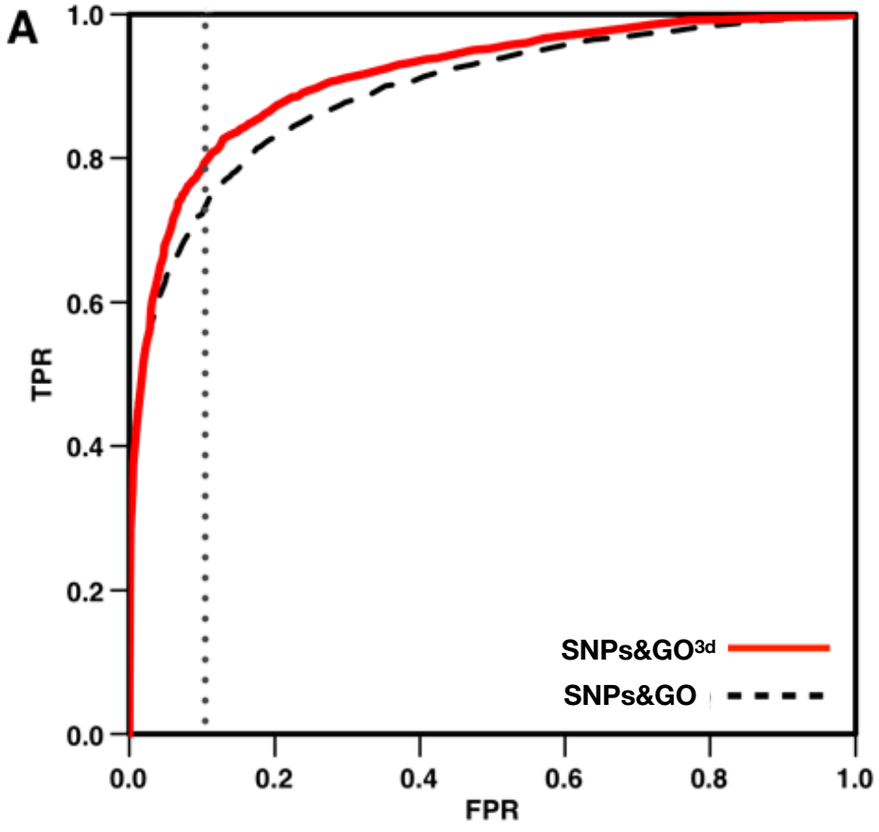


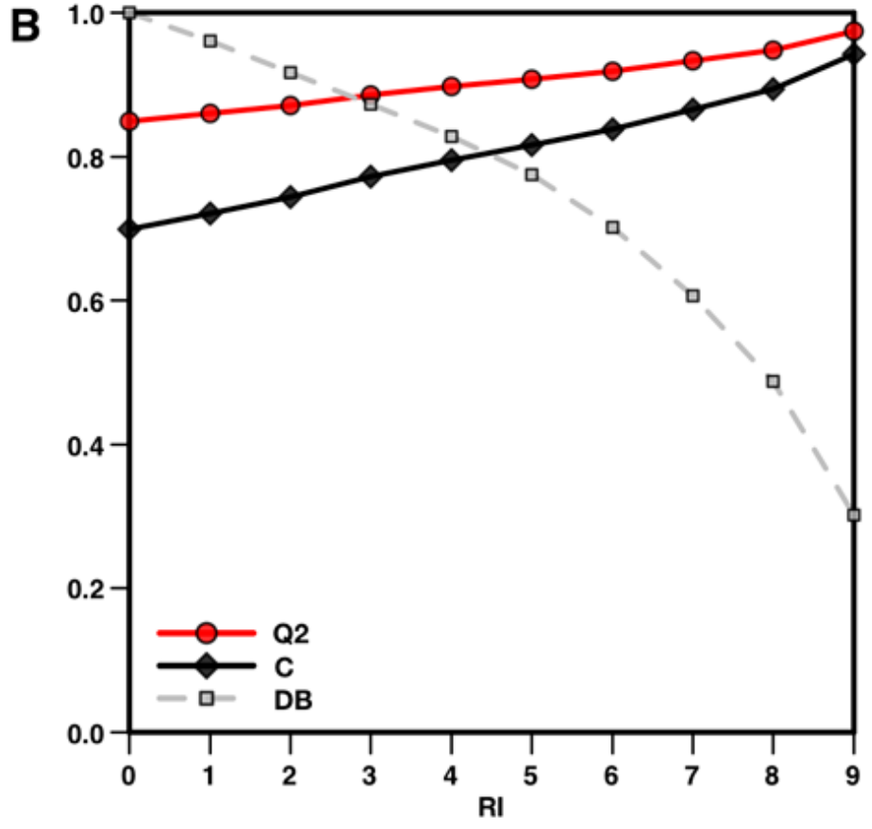


Sequence vs Structure

The structure-based method results in better accuracy with respect to the sequence-based one. Structure based prediction are 3% more accurate and correlation coefficient increases of 0.06. If 10% of FP are accepted the TPR increases of 7%.

	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC
SNPs&GO	0.82	0.81	0.83	0.82	0.81	0.64	0.89
SNPs&GO^{3d}	0.85	0.84	0.87	0.86	0.83	0.70	0.92

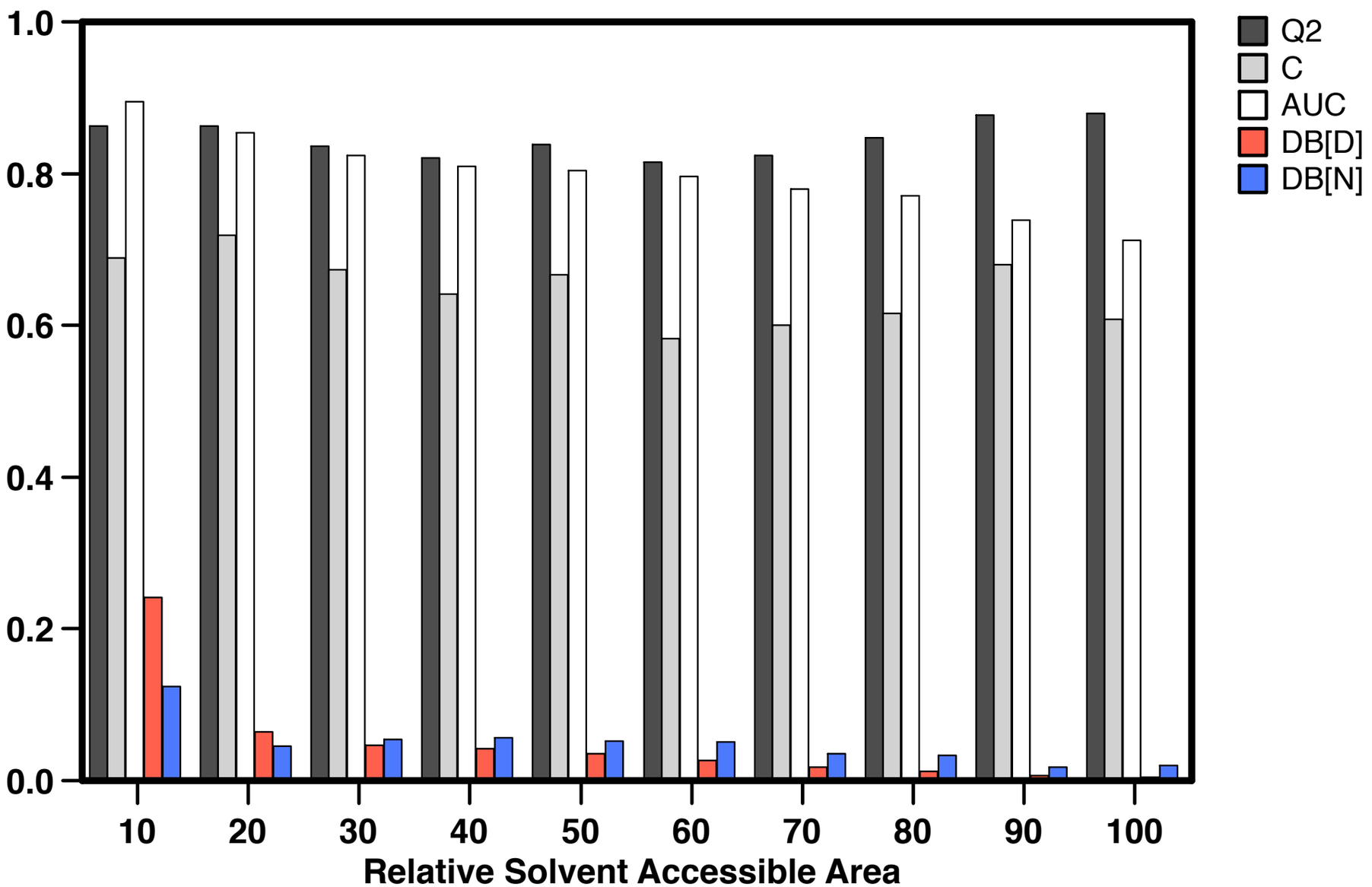




http://snps.biofold.org/snps-and-go

Accuracy vs Accessibility

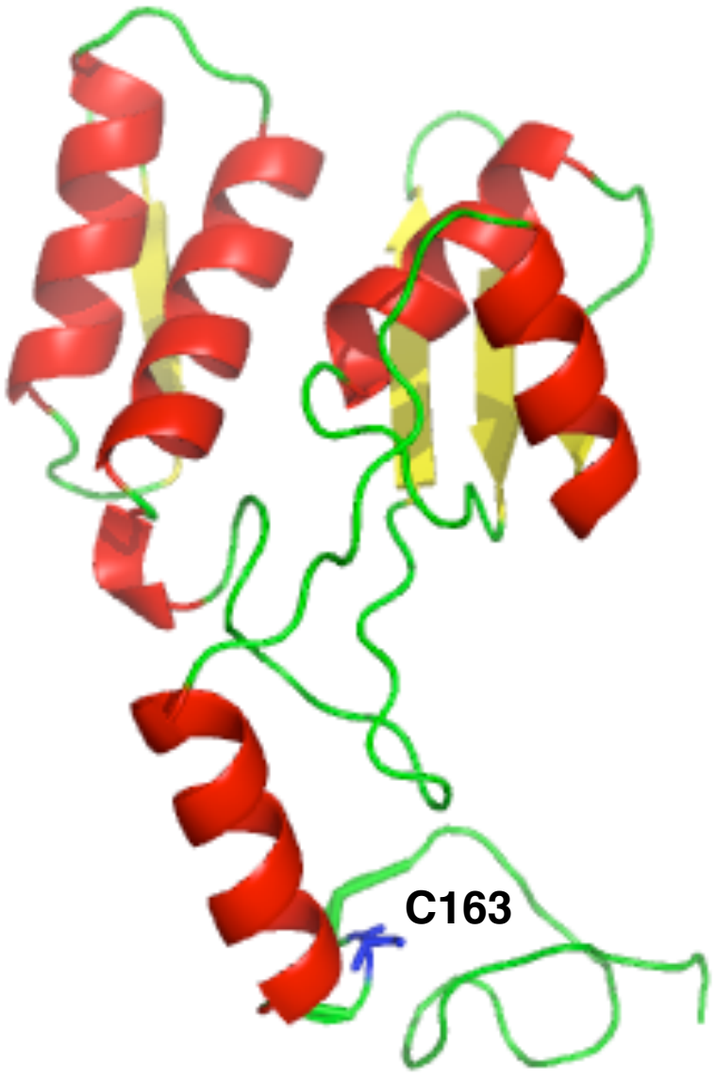
The predictions are more accurate for mutations occurring in buried region (0-30%). Mutations of exposed residues results in lower accuracy.

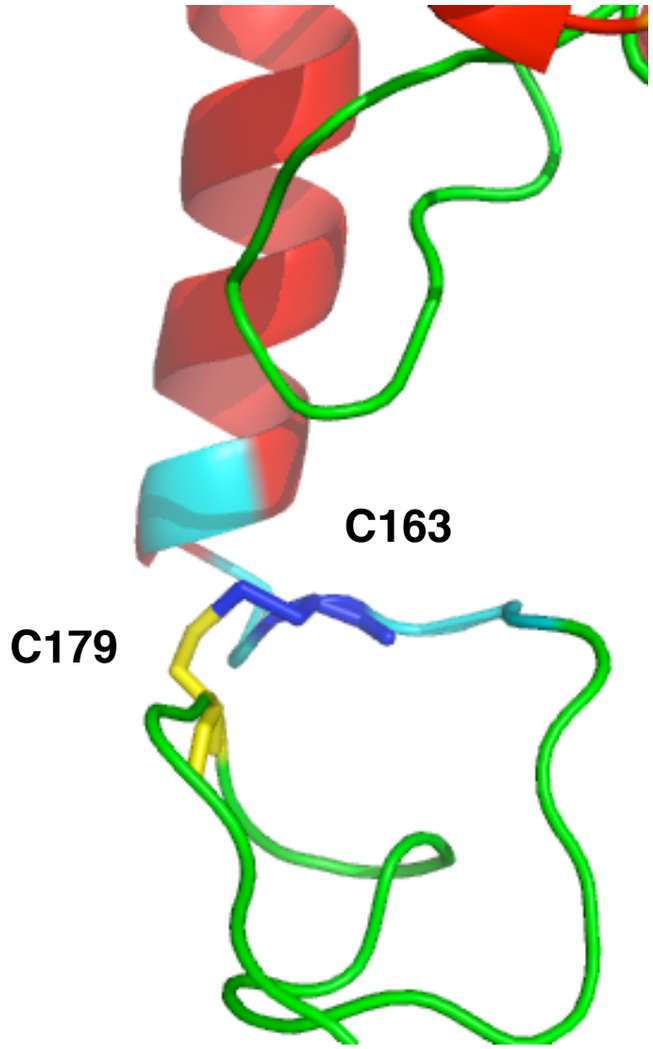


Prediction example

Damaging missing Cys-Cys interaction in the Glycosylasparaginase. The mutation p.Cys163Ser results in the loss of the disulfide bridge between Cys163 and Cys179. This SAP is responsible for Aspartylglucosaminuria.

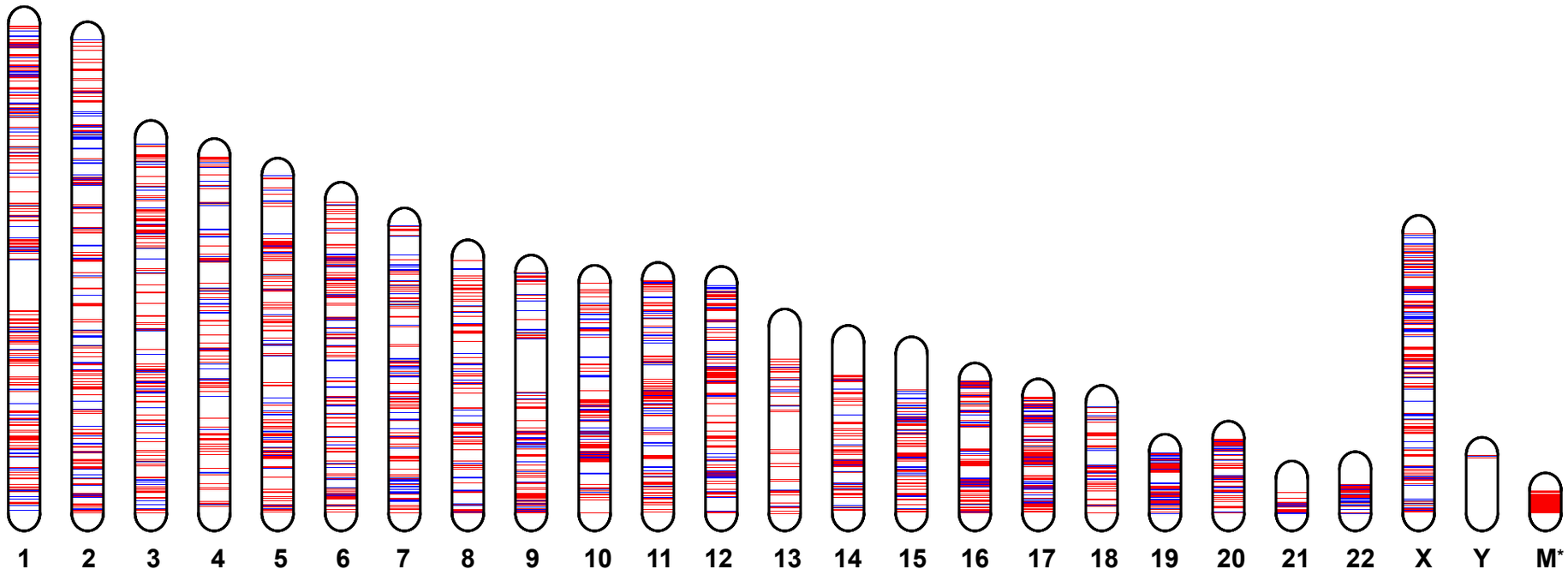
1APY: Chain A, Res: 2.0 Å





Whole-genome predictions

Most of the genetic variants occur in non-coding region that represents >98% of the whole genome.

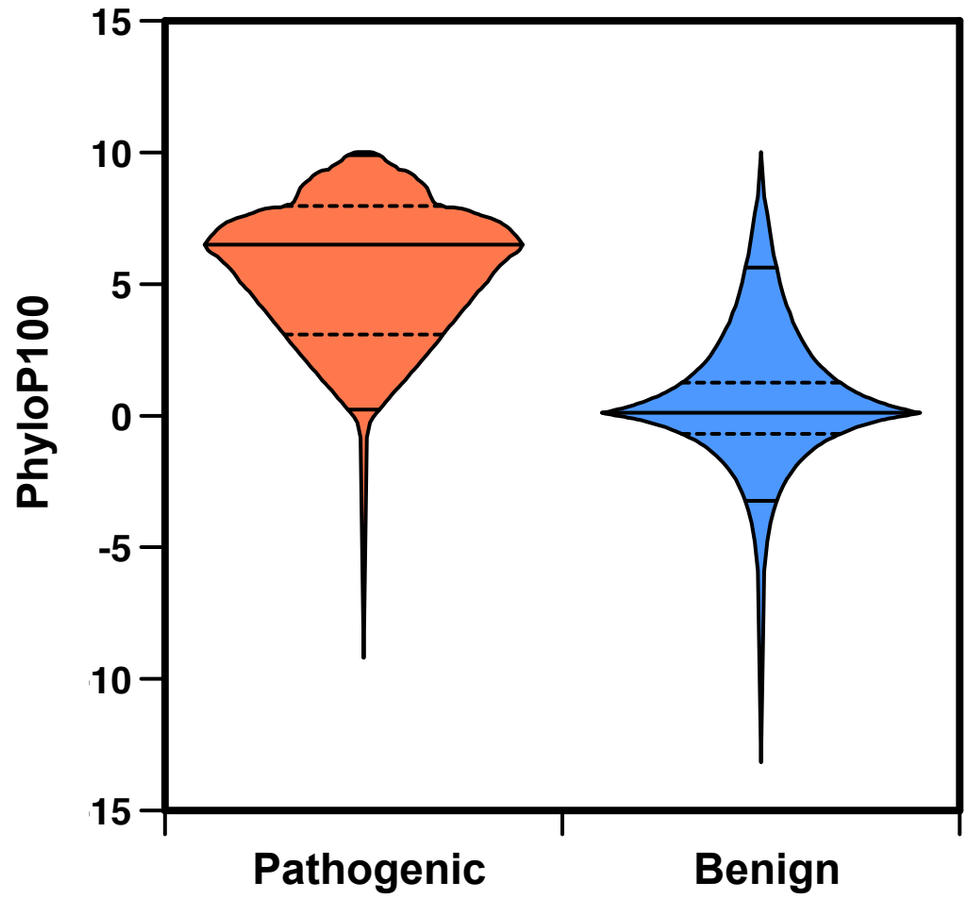


Predict the effect of SNVs in non-coding region is a challenging task because conservation is more difficult to estimate.

Sequence alignment is more complicated for sequences from non-coding regions.

PhyloP100 score

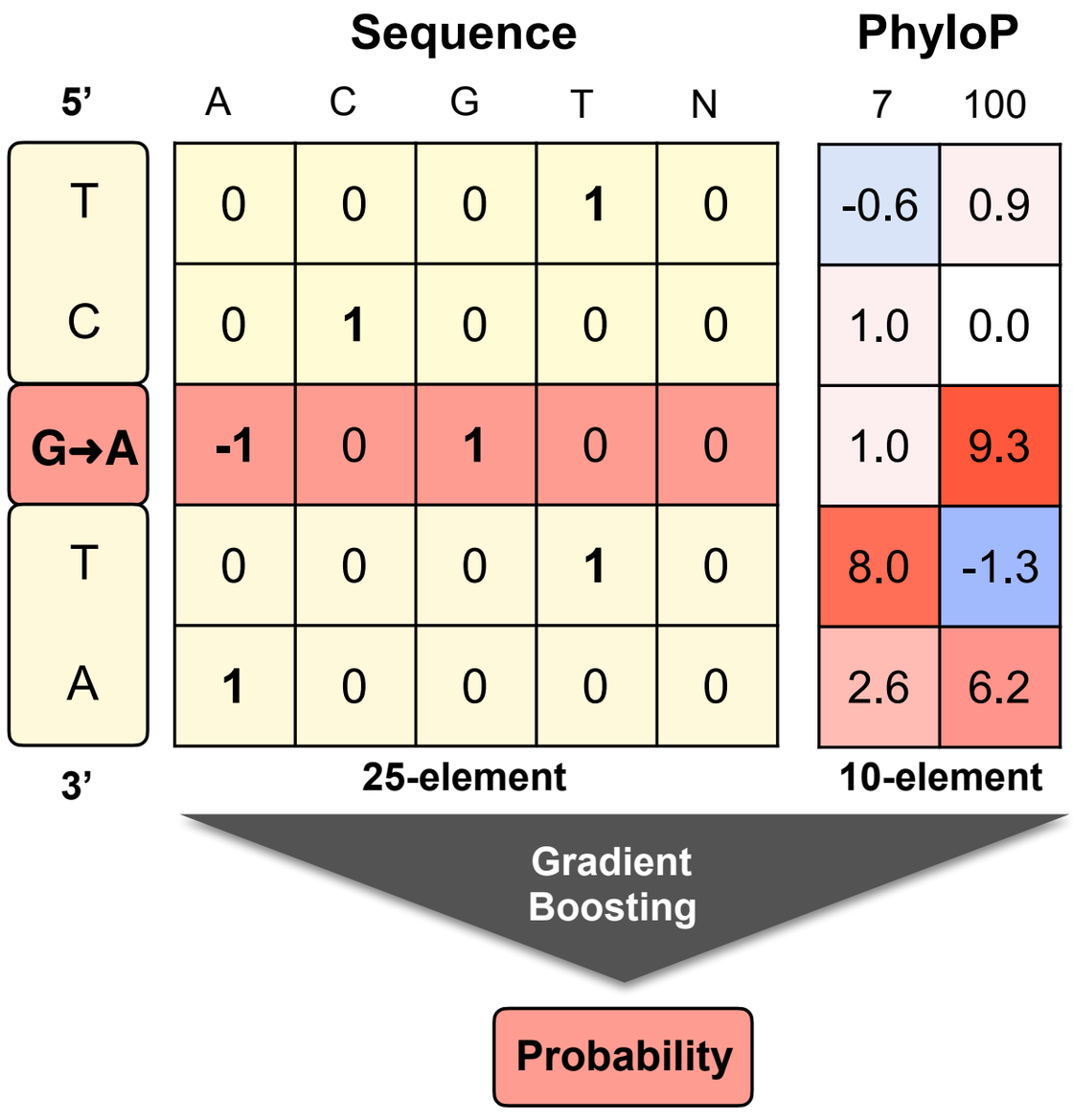
Conservation analysis based on the pre-calculated score available at the UCSC revealed a significant difference between the distribution of the PhyloP100 scores in Pathogenic and Benign SNVs.



PhyloP: Pollard et al,, Genome Research 2010

PhD-SNPg

PhD-SNPg is a simple method that takes in input 35 sequence-based features from a window of 5 nucleotides around the mutated position.

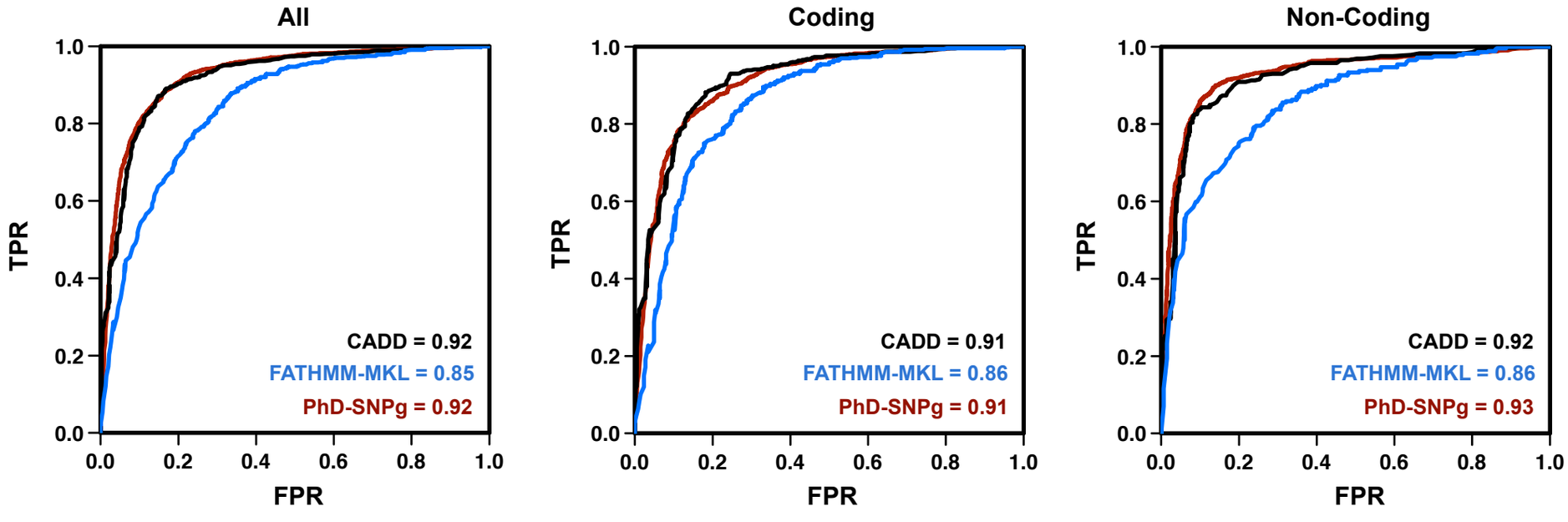


http://snps.biofold.org/phd-snpg/

Benchmarking

PhD-SNP^{g} has been tested in cross-validation on a set of 35,802 SNVs and on a blind set of 1,408 variants recently annotated.

	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhD-SNP^{g}	0.861	0.774	0.884	0.925	0.847	0.715	0.884	0.924
Coding	0.849	0.671	0.845	0.938	0.850	0.651	0.892	0.908
Non-Coding	0.876	0.855	0.911	0.901	0.839	0.753	0.869	0.930



Capriotti and Fariselli. (2017) Nucleic Acids Res. PMID: 28482034.

Blind Validation

CAGI experiments

The Critical Assessment of Genome Interpretation is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation.

https://genomeinterpretation.org/

The P16 challenge

CDKN2A is the most common, high penetrance, susceptibility gene identified to date in familial malignant melanoma. $p16^{INK4A}$ is one of the two oncosuppressor which promotes cell cycle arrest by inhibiting cyclin dependent kinase (CDK4/6).

Challenge: Evaluate how different variants of p16 protein impact its ability to block cell proliferation.

Provide a number between 50% that represent the normal proliferation rate of control cells and 100% the maximum proliferation rate in case cells.

SNPs&GO prediction

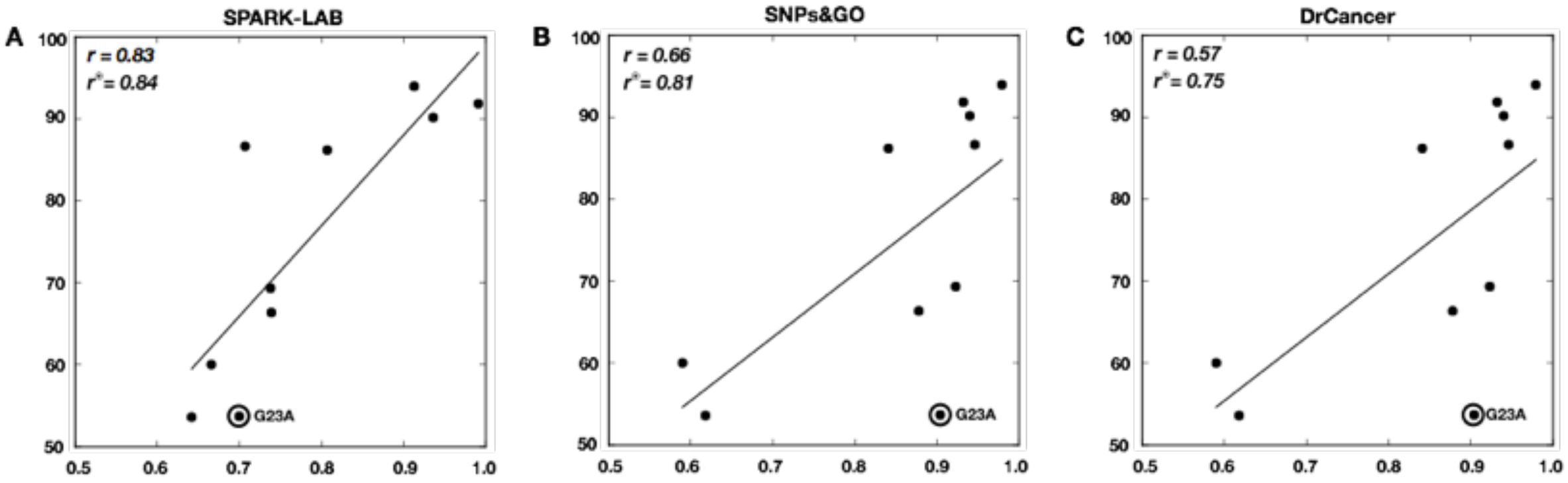
Proliferation rates predicted using the output of SNPs&GO without any optimization.

Variant	Prediction	Real	Δ	%WT	%MUT
G23R	0.932	0.918	0.014	84	0
G23S	0.923	0.693	0.230	84	1
G23V	0.940	0.901	0.039	84	0
G23A	0.904	0.537	0.367	84	2
G23C	0.946	0.866	0.080	84	0
G35E	0.590	0.600	0.010	12	14
G35W	0.841	0.862	0.021	12	0
G35R	0.618	0.537	0.081	12	4
L65P	0.878	0.664	0.214	15	1
L94P	0.979	0.939	0.040	56	0

P16 predictions

SNPs&GO resulted among the best methods for predicting the impact of P16INK4A variants on cell proliferation.

Method	Q2	AUC	MC	RMSE	$r_{Pearson}$	$r_{Spearman}$	$r_{KendallTau}$
SPARK-LAB	0.900	0.920	0.816	0.30	0.595	0.619	0.443
SNPs&GO	0.700	0.880	0.500	0.33	0.575	0.616	0.445
DrCancer	0.600	0.840	0.333	0.46	0.477	0.495	0.409



Capriotti et al. (2017) Human Mutations. PMID: 28102005.

The NAGLU challenge

NAGLU is a lysosomal glycohydrolyase which deficiency causes a rare disorder referred as Sanfilippo B disease

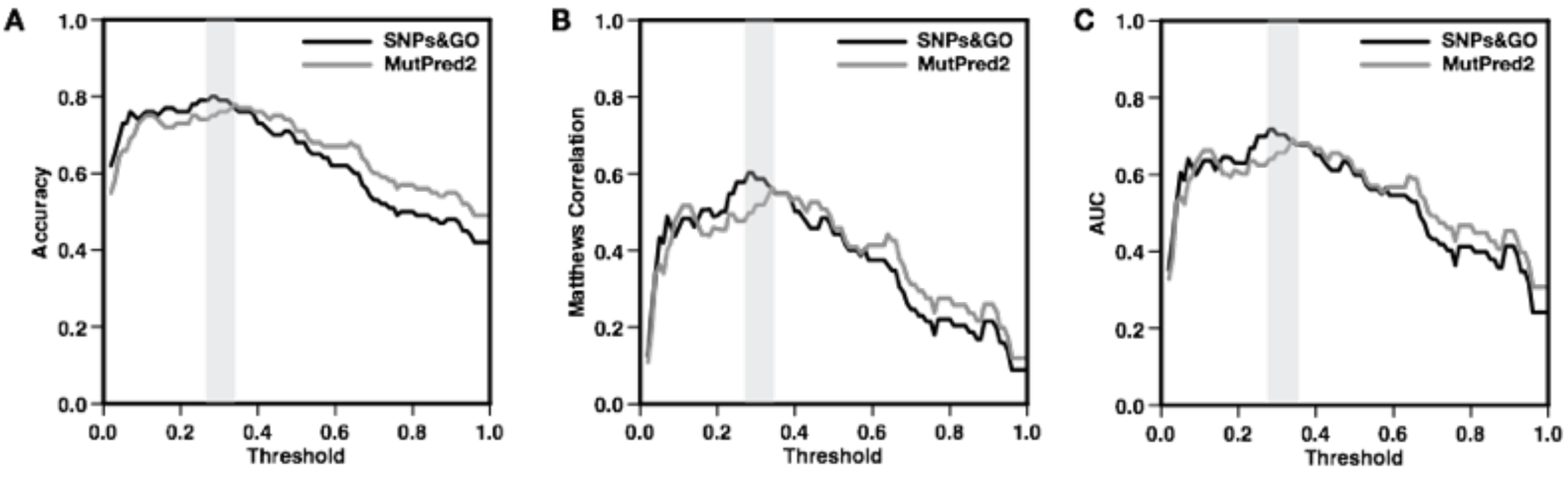
Challenge: Predict the effect of the 165 variants on NAGLU enzymatic activity.

The submitted prediction should be a numeric value ranging from 0 (no activity) to 1 (wild-type level of activity).

A posteriori evaluation

I performed a posteriori evaluation of the performance based on my version of the predictor and found that SNPs&GO reaches similar accuracy than the best method (MutPred2)

Method	Q2	AUC	MC	RMSE	$r_{Pearson}$	$r_{Spearman}$	$r_{KendallTau}$
MutPred2	0.780	0.850	0.565	0.30	0.595	0.619	0.443
SNPs&GO	0.800	0.854	0.603	0.33	0.575	0.616	0.445
SNPs&GO^{09}	0.750	0.749	0.499	0.46	0.477	0.495	0.409



Conclusions

- The machine learning methods based on sequence and structural information, trained to predict the sign and the value of ΔΔG, reach a good level of accuracy.

- Evolutionary information are important for predicting deleterious variants. Wild-type residues in disease-related sites are more conserved than in neutral sites.

- Protein structure information improves performance of machine learning methods to discriminate between disease-causing and neutral variants.

- Nucleotide conservation is an important feature to predict the impact of SNVs in non coding regions

Acknowledgments

Structural Genomics @CNAG
Marc A. Marti-Renom
Francois Serra

Computational Biology and Bioinformatics Research Group (UIB)
Jairo Rocha

Division of Informatics at UAB
Malay Basu
Division Clinical Immunology & Rheumatology
Harry Schroeder
Mohamed Khass

Helix Group (Stanford University)
Russ B. Altman
Jennifer Lahti
Tianyun Liu
Grace Tang

Bologna Biocomputing Group
Rita Casadio
Pier Luigi Martelli
University of Torino
Piero Fariselli
University of Camerino
Mario Compiani

Mathematical Modeling of Biological Systems (University of Düsseldorf)
Markus Kollmann
Linlin Zhao

Other Collaborations
Yana Bromberg, Rutger University, NJ
Hannah Carter, UCSD, CA
Francisco Melo, Universidad Catolica, Chile
Sean Mooney, Buck Institute, Novato
Cedric Notredame, CRG Barcelona
Gustavo Parisi, Univesidad de Quilmes
Frederic Rousseau, KU Leuven
Joost Schymkowitz, KU Leuven

FUNDING
NIH: 1R21 AI134027- 01A1
MIUR: FFABR
UNIBO: International Cooperation
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/