The use of evolutionary information improves the prediction of disease related protein mutations.

Emidio Capriotti[1], Leonardo Arbiza[3], Rita Casadio[4], Joaquín Dopazo[2], Hernán Dopazo[3] and Marc A. Marti-Renom[1]

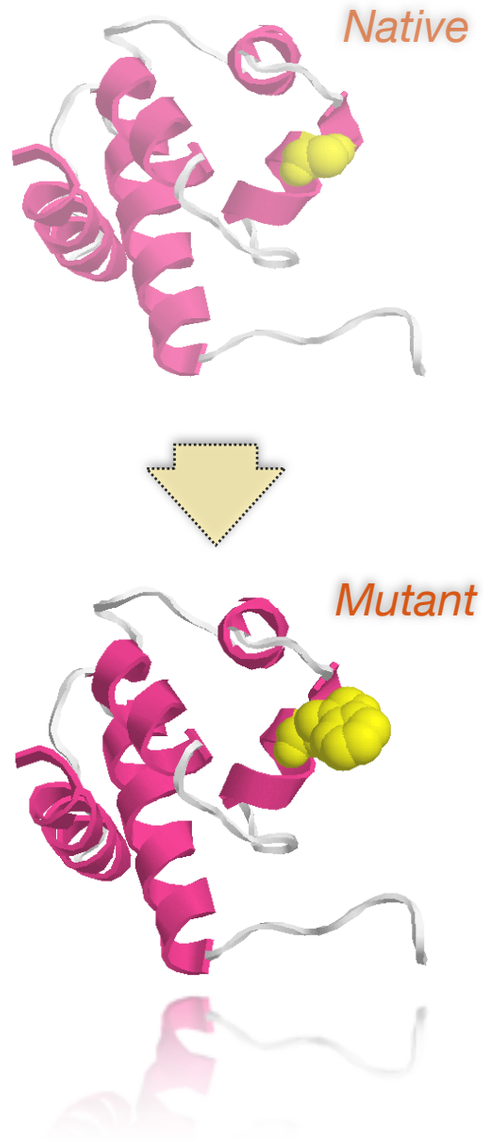
Structural Genomics Unit[1]
Pharmacogenomicsand Comparative Genomics Unit[2]
Functional Genomics Unit[3]
Bioinformatics Department
Prince Felipe Resarch Center (CIPF), Valencia, Spain
http://sgu.bioinfo.cipf.es
http://bioinfo.cipf.es

Laboratory of Biocomputing[4]
Department of Biology
University of Bologna Bologna, Italy
http://www.biocomp.unibo.it

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Summary

- Introduction
- Mutation and Disease
  - Problem definition
  - SNP Databases
  - Datasets
- Methods
  - SVM-based methods
  - Selective pressure
  - Codon-based information methods
  - Results
- Conclusions



Single Nucleotide Polymorphism

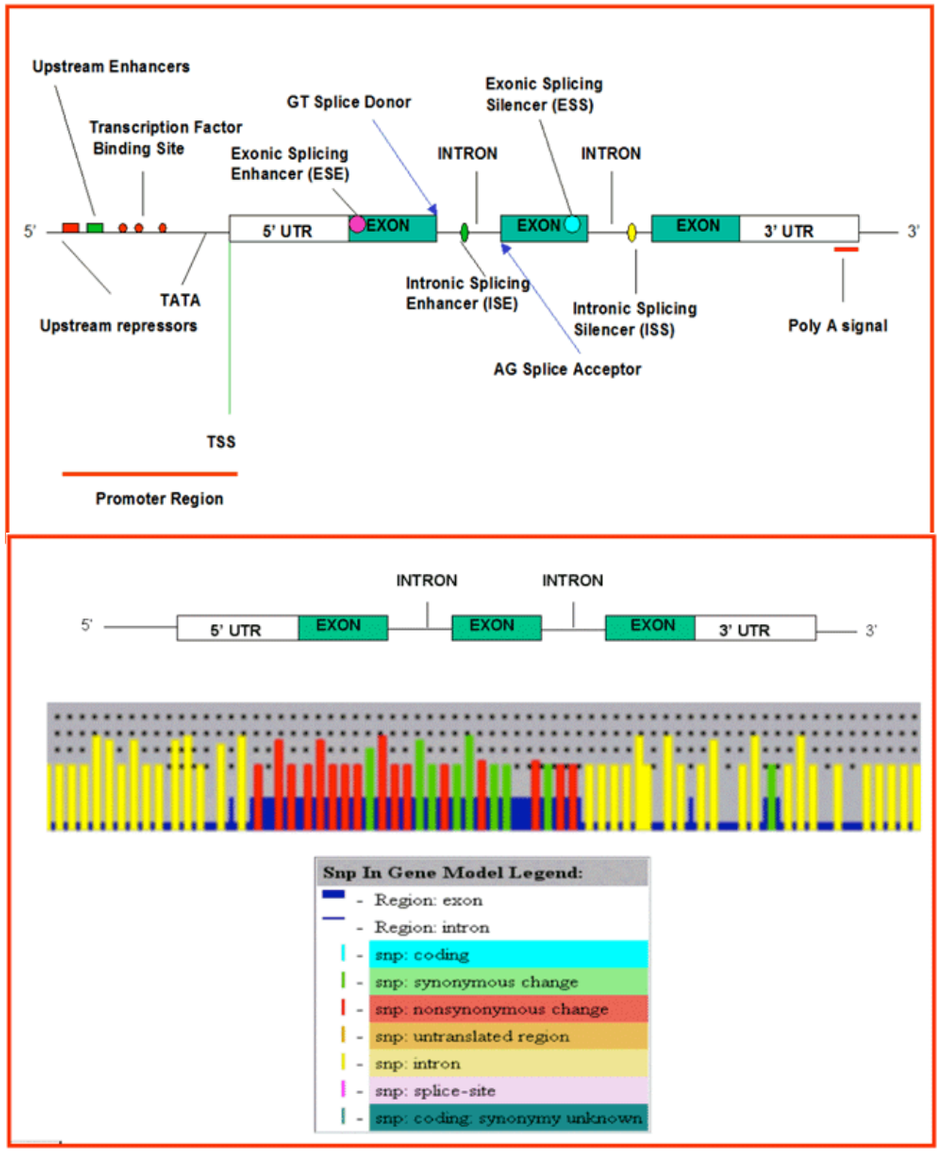
Single Nucleotide Polymorphism or SNP
is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of the species.
Usually one will want to refer to SNPs when the population frequency is ≥ 1%

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

Coding SNPs can be subdivided into two groups:

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

Non-synonymous: when single base substitutions cause a change in the resultant amino acid.



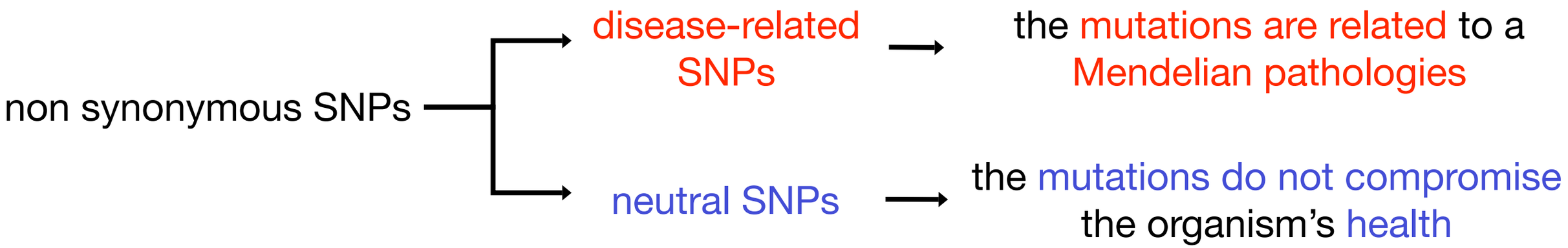
http://www.ncbi.nlm.nih.gov

SNPs and disease

Single nucleotide polymorphisms are the most common type of genetic variations in human accounting for about 90% of sequence differences (Collins et al., 1998).

Studying SNPs distribution in different human populations can lead to important considerations about the history of our species (Barbujani and Goldstein, 2004; Edmonds et al., 2004).

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

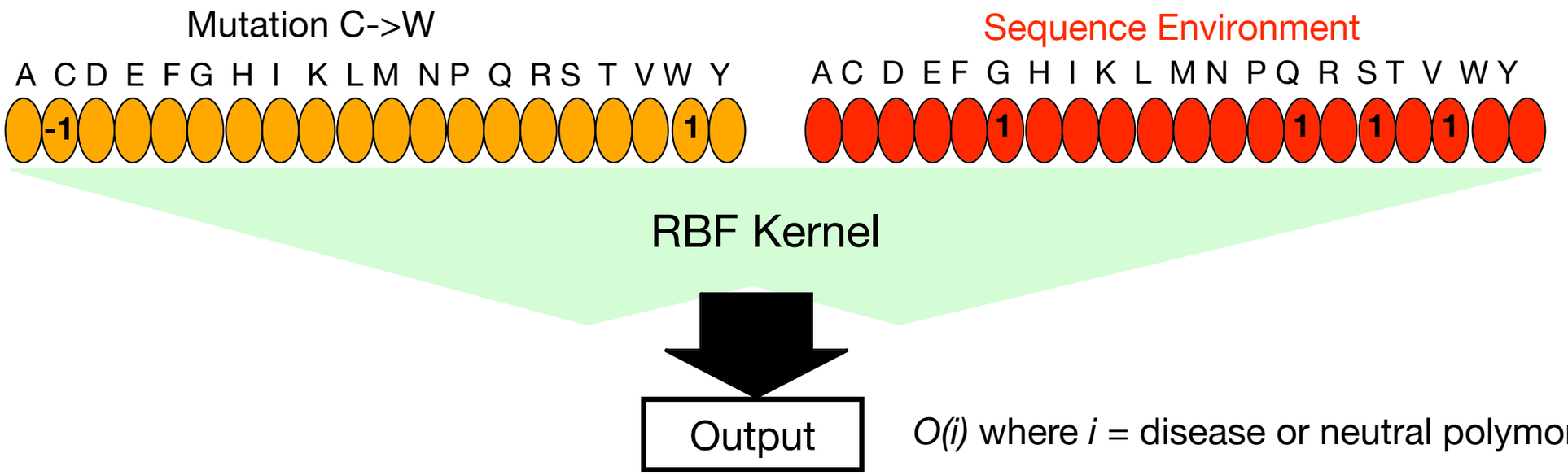


SNP dataset

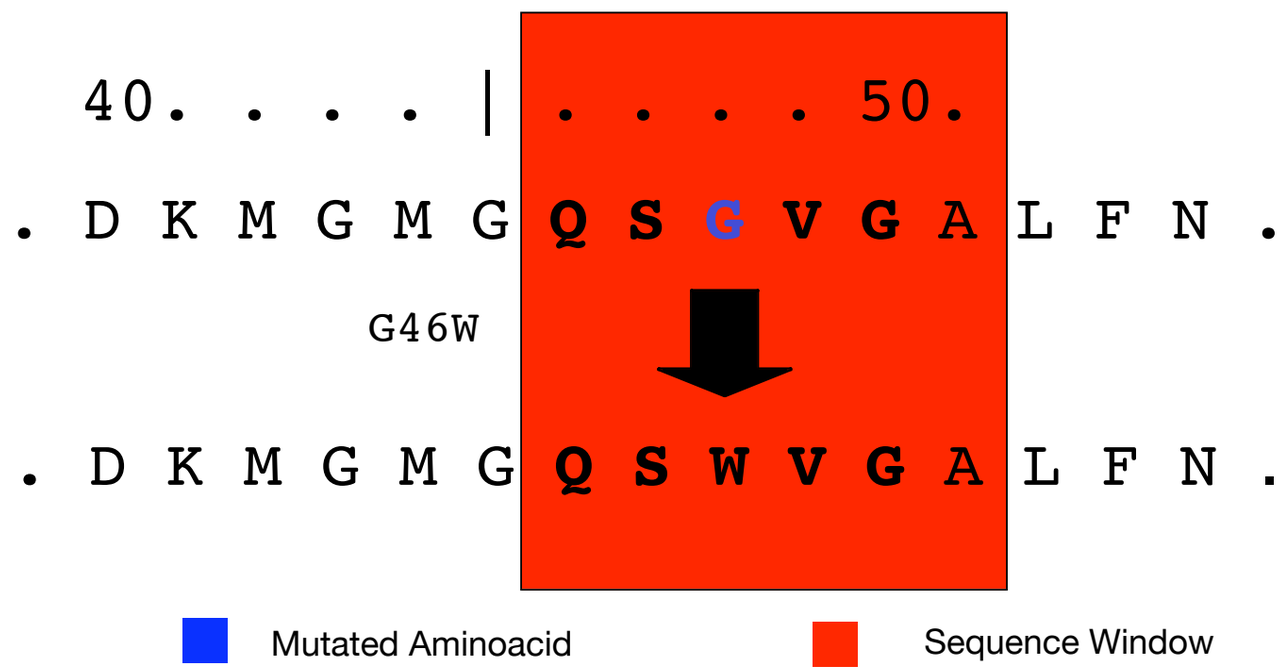
	Mutation	Disease	Neutral	Proteins
Single point mutation with reported effect	**21,185**	12,944	8,241	3,587
Single point mutation with reported effect and profile	**8,718**	3,852	4,866	2,538

from SwissProt (Dec 2005)

Sequence-based predictor

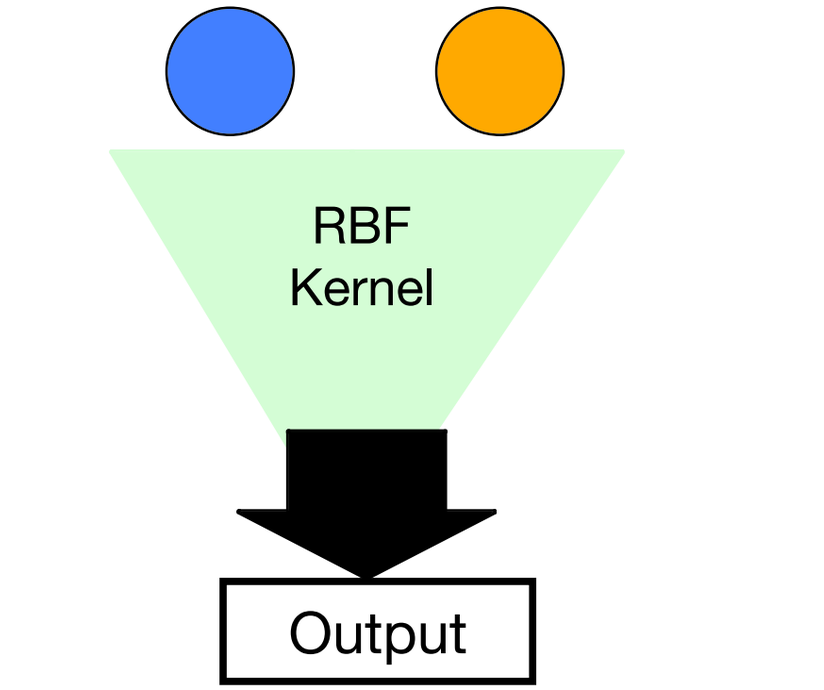


SVM-SEQUENCE: 20 element vector that describes the aminoacid mutation,
20 more input neurons (40 in total) encoding the sequence residue environment



Profile-based predictor

Evolutionary Information derived from sequence profiles are important for detecting mutations that affect human health.



O(i) where *i* = disease or neutral polymorphism

MSA

1	Y	K	D	Y	H	S	-	D	K	K	K	G	E	L	-	-
2	Y	R	D	Y	Q	T	-	D	Q	K	K	G	D	L	-	-
3	Y	R	D	Y	Q	S	-	D	H	K	K	G	E	L	-	-
4	Y	R	D	Y	V	S	-	D	H	K	K	G	E	L	-	-
5	Y	R	D	Y	Q	F	-	D	Q	K	K	G	S	L	-	-
6	Y	K	D	Y	N	T	-	H	Q	K	K	N	E	S	-	-
7	Y	R	D	Y	Q	T	-	D	H	K	K	A	D	L	-	-
8	G	Y	G	F	G	-	-	L	I	K	N	T	E	T	T	K
9	T	K	G	Y	G	F	G	L	I	K	N	T	E	T	T	K
10	T	K	G	Y	G	F	G	L	I	K	N	T	E	T	T	K

sequence position

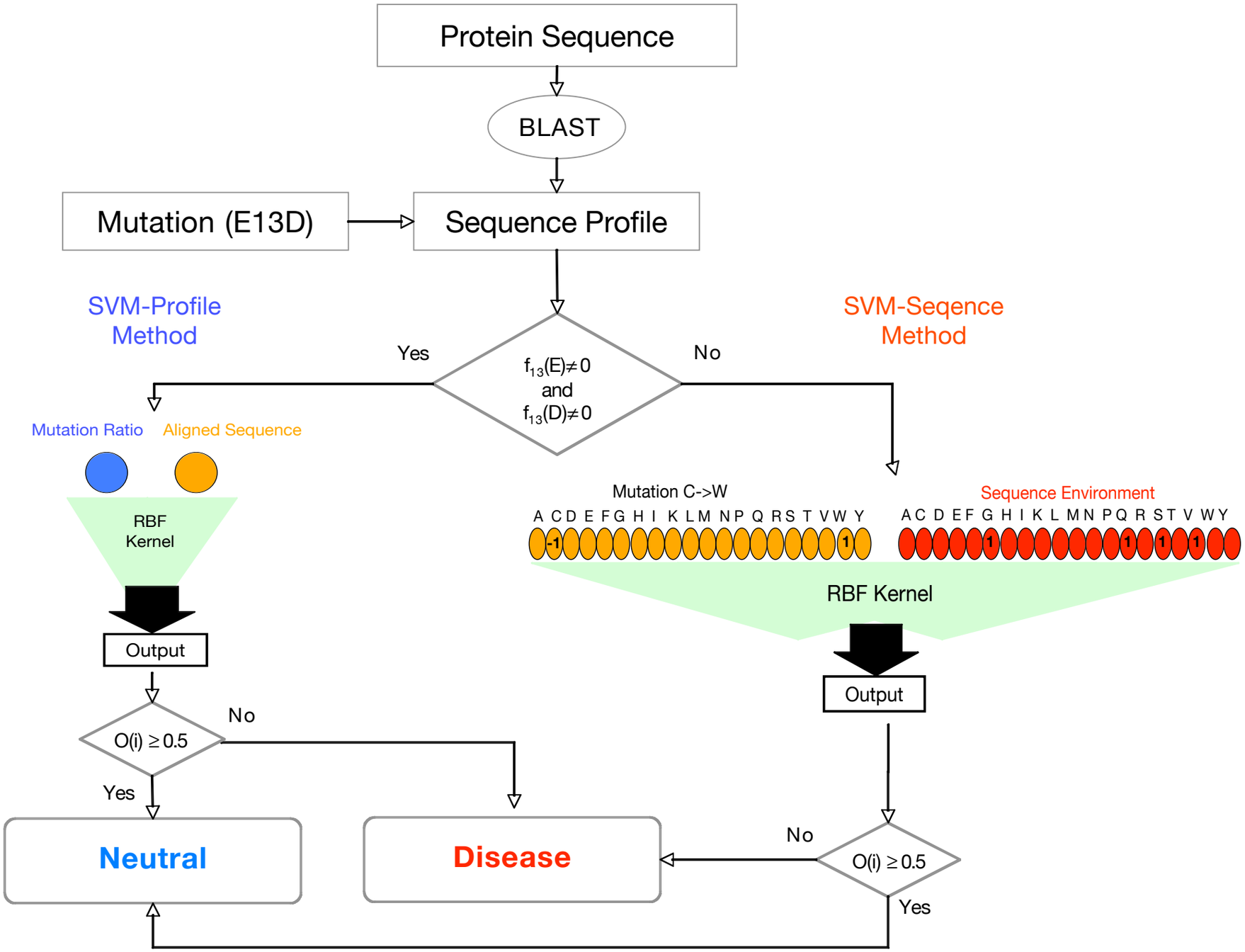
Sequence profile

A	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	70	0	0	0	0	60	0	0	0	0	20	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	70	0	0	0
F	0	0	0	10	0	33	0	0	0	0	0	0	0	0	0	0
G	10	0	30	0	30	0	100	0	0	0	0	50	0	0	0	0
H	0	0	0	0	10	0	0	10	30	0	0	0	0	0	0	0
K	0	40	0	0	0	0	0	0	10	100	70	0	0	0	0	100
I	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0
N	0	0	0	0	10	0	0	0	0	0	30	10	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	40	0	0	0	30	0	0	0	0	0	0	0
R	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	33	0	0	0	0	0	0	10	10	0	0
T	20	0	0	0	0	33	0	0	0	0	0	30	0	30	100	0
V	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0
W	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	70	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0

SVM-PROFILE: Aligned Sequence: number of aligned sequences in the mutated position
Mutation Ratio: ratio between the frequencies in the sequence profile of wild-type versus mutated residues in the considered position.

Hybrid method structure

Hybrid Method is based on a decision tree with SVM-Sequence coupled to SVM-Profile



Classification results

SVM –Sequence is more accurate in the prediction of disease related mutations and SVM-Profile is more accurate in the prediction of neutral polymorphism.
The two methods have the same Q2 level.

	Q2	P[D]	Q[D]	P[N]	Q[N]	C
SVM-Sequence	0.70	0.71	0.84	0.65	0.46	0.34
SVM-Profile	0.70	0.74	0.49	0.68	0.86	0.39
HybridMeth	0.74	0.80	0.76	0.65	0.70	0.46

D = Disease related N = Neutral

The Hybrid Method have higher accuracy than the previous two methods increasing the accuracy up to 74% and the correlation coefficient up to 0.46.

Comparison with other predictors

Hybrid method overcomes in accuracy and correlation the other available methods and provides all the required predictions (see column %PM).

Hybrid method shows a larger value of accuracy with respect to SIFT and although the quality of HybridMeth is comparable to PolyPhen our method needs less information

Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	%PM
PolyPhen	0.72	0.62	0.72	0.80	0.73	0.44	93
SIFT	0.67	0.76	0.67	0.56	0.66	0.33	94
HybridMeth	0.74	0.80	0.76	0.65	0.70	0.46	100

D = Disease related N = Neutral

http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi

Capriotti et al. (2006) *Bioinformatics*, 22; 2729-2734.

Selective pressure

Comparison of relative rates of synonymous (silent) and non-synonymous (amino acid-altering) mutations provides a means for understanding the mechanisms of molecular sequence evolution.

The non-synonymous/synonymous mutation rate ratio ω

$$\omega = \frac{dN}{dS}$$

is an important indicator of selective pressure at the protein level:

ω = 1 meaning neutral selection.
ω < 1 purifying selection.
ω > 1 positive selection.

Dataset

From the dataset used in the previous work we selected only mutation for which it is possible to evaluate the selective pressure.

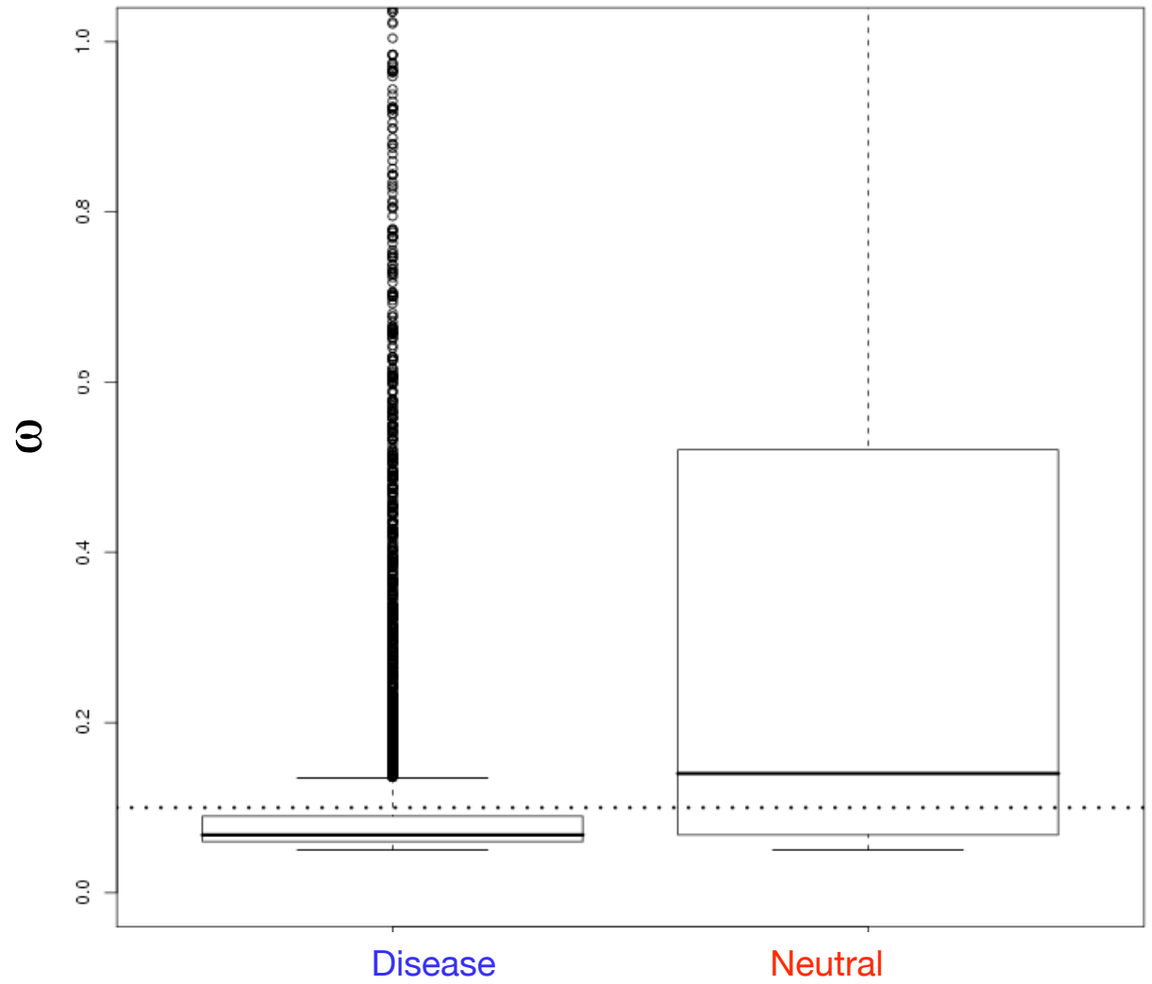
	Dataset	Mutation	Disease	Neutral	Proteins
Single point mutation with reported effect	DBSEQ*	21,185	12,944	8,241	3,587
Single point mutation with evaluable ω	**HM-Dec05***	**8,987**	6,220	2,767	1,434
Single point mutation of new sequences with evaluable ω	**HM-Dec06****	**2,008**	804	1,204	720

*from SwissProt (Dec 2005)
**from SwissProt (Dec 2006)

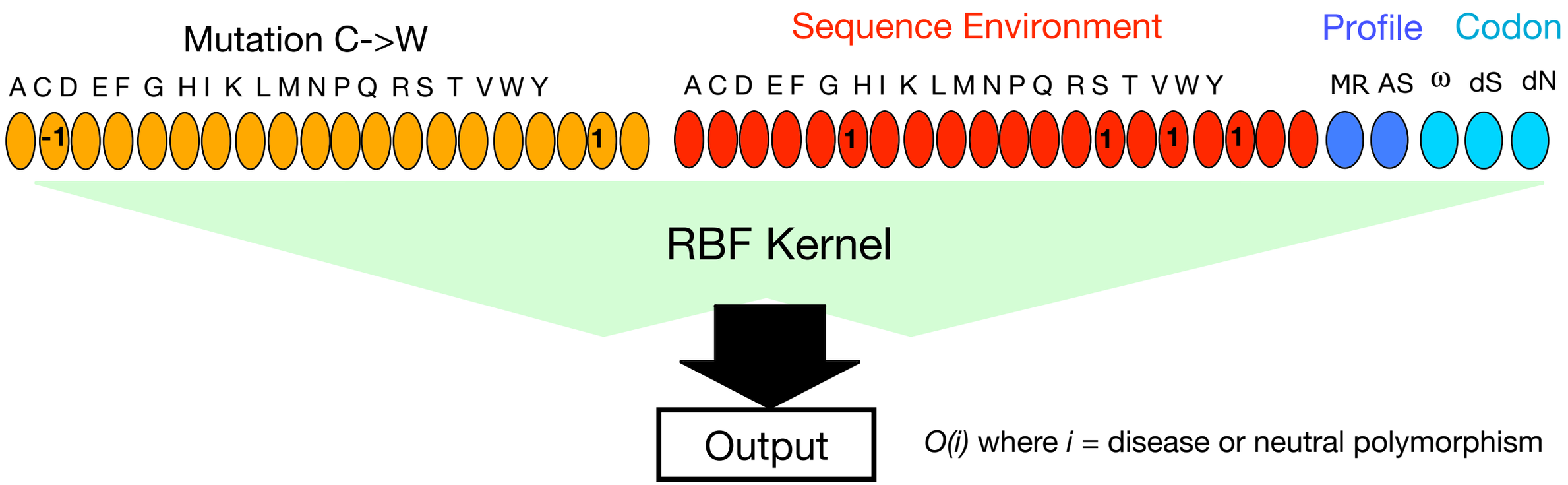
The omega value and disease related mutation

In a previous work performed on 40 human disease genes, has been demonstrated that residues evolving under strong selective pressures (ω<0.1) are significantly associated with human disease (Arbiza et al. (2006) JMB, 358 :1390-1404).

We carried out a similar analysis on the dataset extracted from SwissProt and we found a statistically significant association between high selective pressures and disease in contrast to low selective pressures and neutral polymorphic variants in human.



Sequence and evolutive - based predictors



O(i) where *i* = disease or neutral polymorphism

SEQ: Mutation+ Sequence Environment
SEQPROF: Mutation+ Sequence Environment + Profile
SEQCOD Mutation+ Sequence Environment + Codon
SEQPROFCOD; Mutation+ Sequence Environment + Profile + Codon

Profile: MR and AS sequence profile information
Codon: omega, dS,dN: selective pressure at codon level, synonymous and non-synonymous rate at branch level.

Classification results

SeqCod and SeqProf methods reach the same level of accuracy of about 79% and when the two different types of evolutive information are used the resulting predictor SeqProfCod overcomes the others showing an overall accuracy of 82%

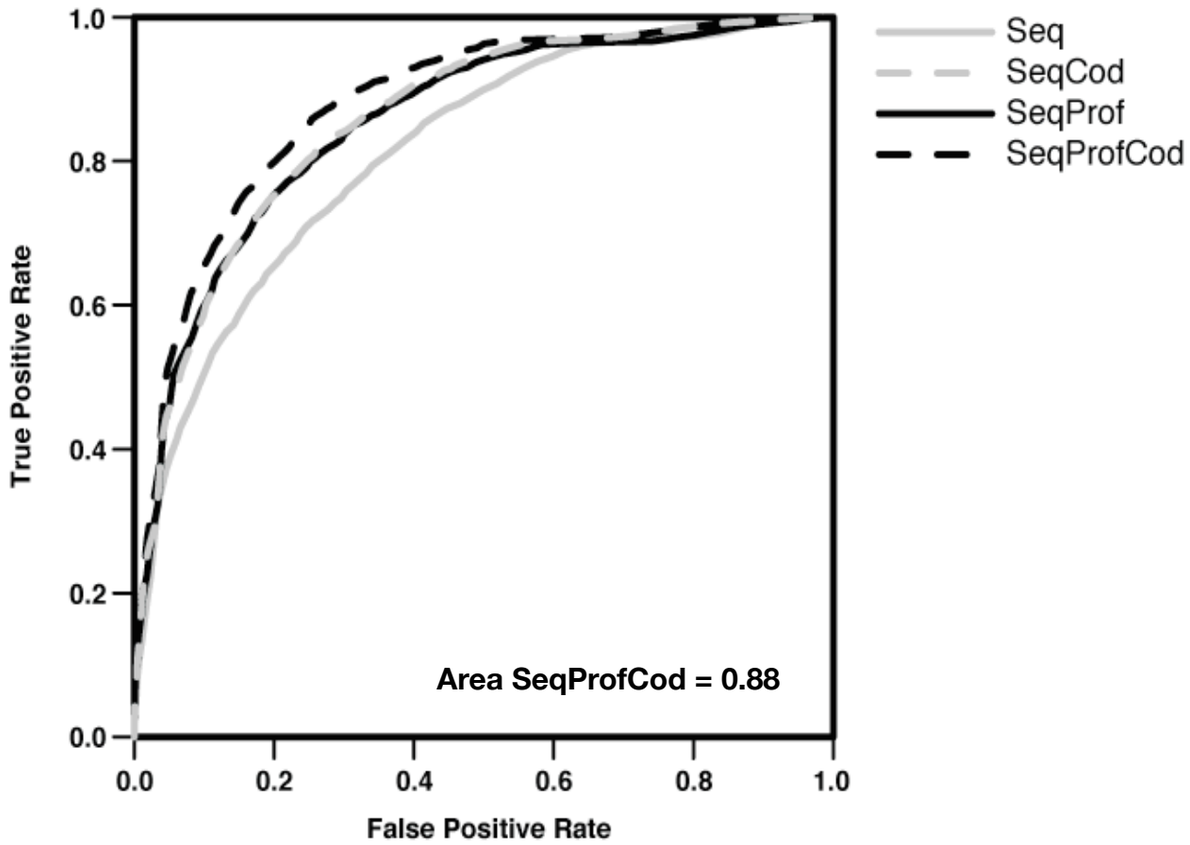
	Q2	P[D]	Q[D]	P[N]	Q[N]	C
Seq	73	86	72	54	74	0.43
SeqCod	79	87	82	64	74	0.53
SeqProf	79	88	81	63	75	0.54
SeqProfCod	82	89	84	68	76	0.59

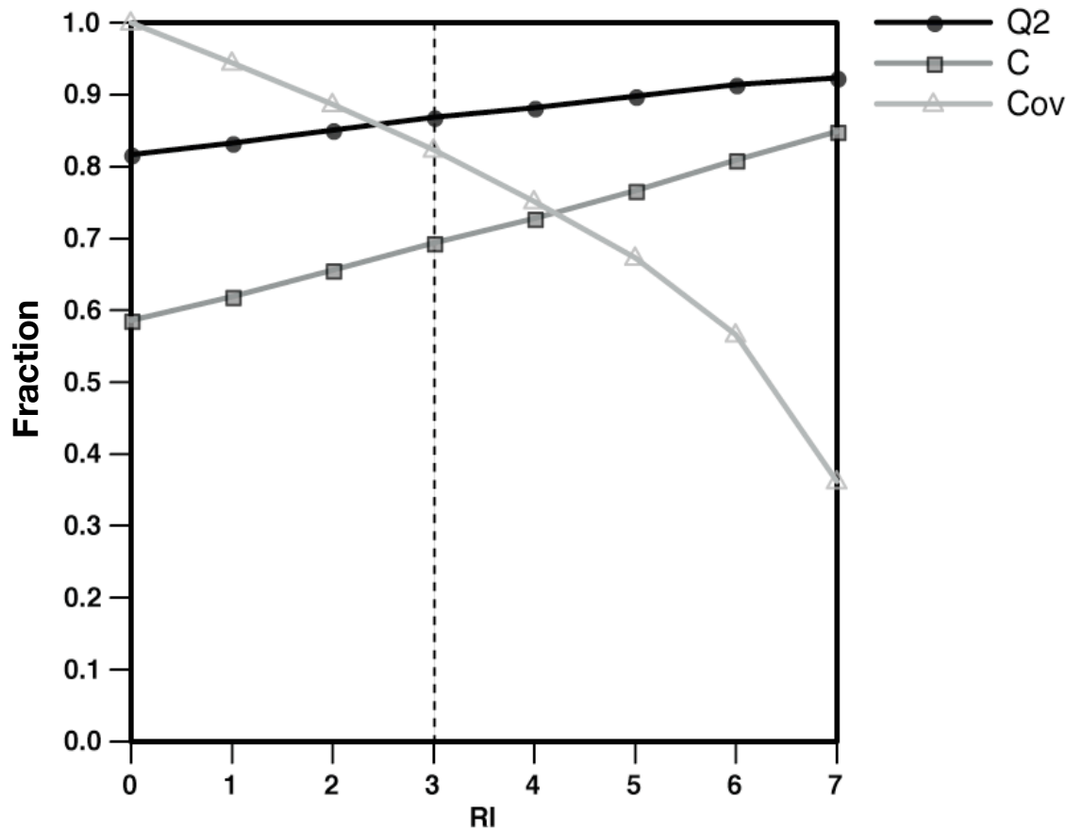
D = Disease related N = Neutral

SeqProfCod method

SeqProfCod has higher accuracy than the previous two methods increasing the accuracy up to 82% and the correlation coefficient to 0.59.

	Q2	P[D]	Q[D]	P[N]	Q[N]	C
SeqProfCod	82	88	84	68	76	0.59





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

Comparison with other predictors

SeqProfCod results in higher accuracy and correlation than the other available methods covering the 100% of the dataset (see column %PM).
SeqProfCod results in higher accuracy with respect to SIFT and although the quality of SeqProf Cod is comparable to PANTHER, when our prediction are selected by RI index the accuracy of our method is higher than PANTHER.

	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
SeqProfCod	82	89	84	68	76	59	100
SIFT	71	84	72	51	69	38	97
PANTHER	74	87	75	53	72	43	83

HM-Dic05: 8987 mutations

	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
SeqProfCod	74	65	79	83	72	48	100
SIFT	71	63	70	78	72	42	96
PANTHER	77	73	71	79	81	52	77

HM-Dic06: 2008 mutations

http://sgu.bioinfo.cipf.es/services/Omidios/

Capriotti et al. (2008) Human Mutation, 29(1):198-204.

Conclusion

Evolutionary information improves the prediction disease related mutation.

A SVM-based method taking into account multiple sequence alignment information reaches good levels of accuracy (Q2=0.79 and C=0.54).

The introduction of codon-based information, such as estimation of selective pressures, improves the accuracy of our predictions (Q2=0.82 and C=0.59).

Although the absolute gains in terms of Q2 appear to be small, the benefits could telescope in a large-scale application such as predicting the effects of the ~83,000 nsSNPs annotated in dbSNP.

Acknowledgments

Structural Genomics Unit (CIPF)
Marc A. Marti-Renom
Emidio Capriotti
Peio Ziarsolo Areitioaurtena

Comparative Genomics Unit (CIPF)
Hernán Dopazo
Leo Arbiza
Francoise Serra

Functional Genomics Unit (CIPF)
Joaquín Dopazo
Fátima Al-Shahrour
José Carbonell
Ignacio Medina
David Montaner
Joaquin Tárraga
Ana Conesa
Toni Gabaldón
Eva Alloza
Lucía Conde
Stefan Goetz
Jaime Huerta Cepas
Marina Marcet
Pablo Minguez
Jordi Burguet Castell
Pablo Escobar

Laboratory of Biocomputing
University of Bologna
Rita Casadio
Remo Calabrese
Piero Fariselli

FUNDING
Prince Felipe Research Center
Marie Curie Reintegration Grant
STREP EU Grant
Generalitat Valenciana
MEC-BIO

http://bioinfo.cipf.es
http://sgu.bioinfo.cipf.es