Selective pressure improves the prediction of disease relate nsSNP in human

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

1. Bioinformatics Department. Centro de Investigación Príncipe Felipe, Valencia, Spain.
2. CIRB. Department of Biology. University of Bologna, Bologna, Italy.

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Aim

Predicting the functional impact of a protein variation is one of the most challenging problems in Bioinformatics. A rapidly growing number of genome-scale studies provide large amounts of experimental data allowing the application of rigorous statistical approaches for predicting if a given single point mutation has or not an impact on human health (1-5). Up until now, existing methods have limited their source data to either protein or gene information. Novel in this work, we take advantage of both and focus on protein evolutionary information by using estimated selective pressures at the codon level.

Methods

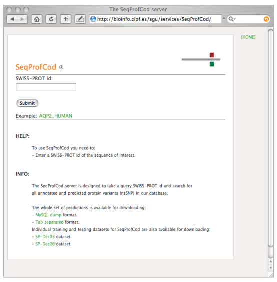
Here we introduce a new method (SeqProfCod) to predict the likeliness that a given protein variant is associated or not with a human disease. We rely on our previous work (4) to extend the implementation of sequence-based and profile-based SVMs to include codon-based estimation of selective pressures at each position of the target sequences. The selective pressure has been calculated evaluating the ratio between non-synonymous to synonymous rates of substitution (ω=dN/dS). The discriminating power of the selective pressure has been discussed in a recent work where we have hypothesized that residues evolving under markedly strong selective pressures ($\omega < 0.1$) are significantly ($p < 0.01$) associated with human disease (5). SeqProfCod combines protein sequence and profile information with selective pressures in a 45 features vector (i.e., 40 features describing the sequence-based information, 2 features describing the profile-based information and 3 features describing the codon-based information).

Results

SeqProfCod has been benchmarked with a large dataset of 8,987 single point mutations from 1,434 human proteins from SWISS-PROT. It achieves 82% overall accuracy and a correlation coefficient of 0.59 indicating that the estimation of the selective pressure helps in predicting the functional impact of single-point mutations. Thus, this study demonstrates the synergic effect of combining the two classical sources of information for predicting the functional effects of protein variants: protein sequence/profile-based information and selective pressures at codon level.

Availability

The results of large-scale application of SeqProfCod over all annotated point mutations in SWISS-PROT are available for download at http://bioinfo.cipf.es/sgu/services/SeqProfCod/.



Example: AQP2_HUMAN

HELP:

To use SeqProfCod you need to:
- Enter a SWISS-PROT id of the sequence of interest.

INFO:

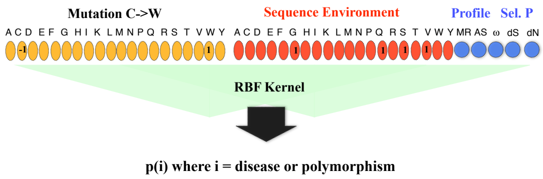
The SeqProfCod server is designed to take a query SWISS-PROT id and search for all annotated and predicted protein variants (nsSNP) in our database.

The whole set of predictions is available for downloading:
- MySQL dump format.
- Tab separated format.
Individual training and testing datasets for SeqProfCod are also available for downloading:
- SP-Dec05 dataset.
- SP-Dec06 dataset.

References

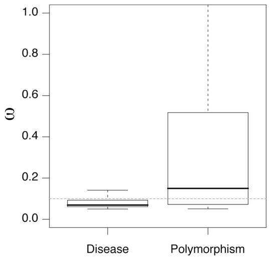
1. V. Ramensky, et al., Nucleic Acids Res 30, 3894 (2002).
2. P. C. Ng and S. Henikoff, NAR 31, 3812 (2003).
3. P. D. Thomas, et al., Nucleic Acids Res 31, 334 (2003).
4. E. Capriotti, et al, Bioinformatics 22, 2729 (2006).
5. L. Arbiza, et al., J Mol Biol 358, 1390 (2006).

SVM training (SeqProfCod)


Mutation C->W
Sequence Environment
Profile
Sel. P
A C D E F G H I K L M N P Q R S T V W Y
A C D E F G H I K L M N P Q R S T V W Y
MR AS ω dS dN
-1
1
1
1
1
1
RBF Kernel
p(i) where i = disease or polymorphism

20 element vector describing the aminoacid mutation
20 element vector describing the sequence residue environment
2 element vector describing the profile information
3 element vector describing the selective pressure at codon level

ω and disease



ω distribution for disease and polymorphism protein variants in the SP-Dec05 dataset.

The boxplot shows the median (horizontal bold line), the upper and lower quartiles (box) and the interquartile range (dashed vertical lines). For visual inspection, a horizontal dotted line indicates value of 0.1.

Classification results

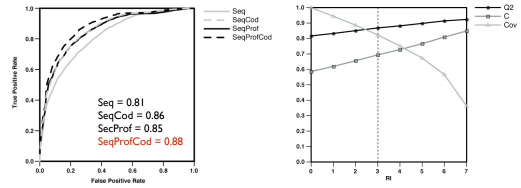
SeqCod and SeqProf methods reach the same level of accuracy of about 0.73 and when the two different types of evolutionary information are used the resulting predictor SeqProfCod reaches a higher accuracy than the others methods (0.82)

	Q2	P[D]	Q[D]	P[N]	Q[N]	C
Seq	0.73	0.86	0.72	0.54	0.74	0.43
SeqCod	0.79	0.87	0.82	0.64	0.74	0.53
SeqProf	0.79	0.88	0.81	0.63	0.75	0.54
SeqProfCod	0.82	0.89	0.84	0.68	0.76	0.59

D = Disease related N = Neutral

Method accuracy

SeqProfCod over-performs the previous two methods increasing the accuracy up to 82% and the correlation coefficient up to 0.59.



SP-Dic05: 8,987 mutations

Comparison with other predictors

SeqProfCod over-performs in accuracy and correlation the other available methods with whole coverage of the data-set (100% PM).

SeqProfCod results in a higher accuracy than SIFT and is comparable to PANTHER
At the same RI index, the accuracy of SeqProfCod is higher than for PANTHER

	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
SeqProfCod	0.82	0.89	0.84	0.68	0.76	0.59	100
SIFT	0.71	0.84	0.72	0.51	0.69	0.38	97
PANTHER	0.74	0.87	0.75	0.53	0.72	0.43	83

SP-Dic05: 8,987 mutations

	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
SeqProfCod	0.74	0.65	0.79	0.83	0.72	0.48	100
SIFT	0.71	0.63	0.70	0.78	0.72	0.42	96
PANTHER	0.77	0.73	0.71	0.79	0.81	0.52	77

SP-Dic06: 2,008 mutations

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