**INPS: predicting the impact of non-synonymous variations on protein stability from sequence**

# INPS: Predicting the Impact of Non-Synonymous Variations on Protein Stability from Sequence.

Piero Fariselli[1,2*], Pier Luigi Martelli[1], Castrense Savojardo[1], Rita Casadio[1]

[1]Biocomputing Group, University of Bologna, Department of Biology, 40126 Bologna

[2]Department of Computer Science and Engineering, University of Bologna, 40127 Bologna, Italy

**ABSTRACT**

**Motivation:** A tool for reliably predicting the impact of variations on protein stability is extremely important for both protein engineering and for understanding the effects of Mendelian and somatic mutations in the genome. Next Generation Sequencing (NGS) studies are constantly increasing the number of protein sequences. Given the huge disproportion between protein sequences and structures, there is a need for tools suited to annotate the effect of mutations starting from protein sequence without relying on the structure.

Here we describe INPS, a novel approach for annotating the effect of non-synonymous mutations on the protein stability from its sequence. INPS is based on a SVM regression and it is trained to predict the thermodynamic free energy change upon single-point variations in protein sequences.

**Results**: We show that INPS performs similarly to the state of the art methods based on protein structure when tested in cross-validation on a non-redundant dataset. INPS performs very well also on a newly generated dataset consisting of a number of variations occurring in the tumor suppressor protein p53. Our results suggest that INPS is a tool suited for computing the effect of non-synonymous polymorphisms on protein stability when the protein structure is not available. We also show that INPS predictions are complementary to those of the state of art, structure-based method mCSM. When the two methods are combined, the overall prediction on the p53 set scores significantly higher than those of the single methods.

**Availability**: The presented method is available as web server at http://inps.biocomp.unibo.it.

## 1 INTRODUCTION

The increasing amount of data generated by the several sequencing initiatives (Hudson et al., 2012; Stratton et al., 2012) calls for accurate and reliable computational approaches to predict the impact of mutations on the phenotype, and possibly for methods to correlate them with diseases (Casadio et al., 2011). More to this, the ability of accurately predicting the impact of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein stability is essential for understanding the effects of human genome variations (Lahti et al., 2012). Several methods have been developed so far to predict the effect of nsSNPs on the protein stability. Some of them require the knowledge of the protein structure: AUTO-MUTE (Masso and Vaisman 2008), CUPSAT (Parthiban et al., 2006), Dmutant (Zhou and Zhou, 2002), FoldX (Guerois et al., 2002), Eris (Yin et al., 2007), PoPMuSiC (Dehouck et al., 2009), SDM (Topham et al., 1997; Worth et al., 2011), mCSM (Pires et al., 2014a) and NeEMO (Giollo et al., 2014). Other methods are based only on protein sequences (iPTREE-STAB: Huang et al., 2007; MuStab: Teng et al., 2010) or can use both protein sequences and protein structures (I-mutant2.0: Capriotti et al., 2005; Imutant3.0 Capriotti et al., 2008; Mupro: Cheng et al., 2006). Besides the single methods, other approaches have been tested, such as meta-predictors (iStable: Chen et al., 2013), filtering approaches based on the available information about mutations in the same protein site (Wainreb et al., 2011) and ensemble predictors (Pires et al., 2014b).

The available methods were trained under different conditions and on different data sets. They address three different questions. Briefly, they can: i) predict the ΔΔG real values (in regression) upon residue substitution, ii) predict whether a residue substitution promotes a ΔΔG increase or decrease (two class predictors), and iii) predict whether a mutation is stabilizing, destabilizing or not affecting the protein stability (three class predictors). Noticeably, it is also very difficult to find a good benchmark test set, since all the methods have to deal with the paucity of the available experimental data. Almost all methods are trained on data derived from the same source: the ProThem database (Kumar et al., 2006).

In 2010, Khan and Vihinen made a thorough evaluation of the different methods considering them as "pure classifiers" (i.e. the methods that predicted the ΔΔG real values were converted into classifiers). The authors showed that the best performing methods (I-Mutant3.0-[structure based], Dmutant, and FoldX) exploit the protein structure information (Khan and Vihinen, 2010).

Recently, Pires et al. (2014a) introduced two relevant advancements when evaluating the performance of different methods. They described: i) a more correct way to avoid similarity between training and testing sets when adopting a cross validation procedure, and ii) a new independent benchmark consisting of a wide range of mutations occurring in the tumor suppressor protein p53, not present in the original ProTherm database (Kumar et al., 2006). In this paper, we take advantage of these efforts, and we describe INPS (a predictor of the Impact of Non-synonymous-variations on Protein Stability), a new method that computes the ΔΔG values of protein variants without requiring the knowledge of the protein structure. We show that when evolutionary information is taken into account, INPS performances are

*To whom correspondence should be addressed.

very close to those obtained by the state-of-the-art methods based on 3D structure, mCSM and Duet. We also show that INPS predictions are complementary to those obtained with mCSM (or Duet) and that their combinations, obtained by averaging the predictions, outperform previously introduced and combined approaches (Pires at al, 2014b).

**Table 1.** Pearson Correlation between S2648 DDG values and sequence-based features.

| Feature | Pearson Correlation | p-value |
|---|---|---|
| BL62 | 0.11 | 1E-8 |
| ΔHy(Hym-Hyw) | 0.28 | 6.0E-49 |
| Mb | 0.17 | 3.0E-19 |
| ΔMW(MWm-MWw) | 0.18 | 3.0E-21 |
| Profile | 0.21 | 1.0E-24 |
| HMM | 0.27 | 5.0E-44 |

w=wild-type residue. m=mutated residue. BL62(w,m)=mutation scored according to the Blosum62 matrix of the substitution wild-type (w) with mutated residue (m). ΔHy(Hym-Hyw)=Hydrophobicity difference between mutated and wild type residues (Kyte and Doolittle, 1982). Mb=mutability value for the wild-type residue (Dayhoff et al., 1978). ΔMW(MWm-MWw)=molecular weight difference between mutated and wild type residues. Profile=difference between the sequence profile positions of the wild-type and mutated residues. HMM= HMM score of the wild-type protein and the mutated protein computed using HMMER program. (Eddy 1998). p-values are computed by means of the Student's t-distribution (Rahman, 1968). See "INPS:input encoding" section for further details.

## 2 MATERIALS AND METHODS

### 2.1 Data Sets

In this paper, we adopted two previously introduced datasets for the prediction of protein stability variations upon single point mutations (Pires et al., 2014a): S2648 and P53. S2648 was originally derived from the ProTherm database (Kumar et al., 2006) and corrected by the authors of the PoPMuSiC algorithm (Dehouck et al., 2009). The data set comprises 2648 single-point variations in 132 different globular proteins. In this paper we adopted the two 5-fold cross-validation procedure introduced by Pires et al., (2014a) and the single training/testing set called "blind" by the authors (Pires et al., 2014a). The first kind of cross-validation fold, labeled as protein-fold ("prot"), groups the variations according to their protein origin (variations belonging to the same protein appear in the same test set). The second type of cross-validation fold, labeled as position-fold ("pos"), groups variations according to their positions along the protein sequence (multiple variations of the same protein position are grouped together in the same test-set). The two different ways of splitting the same non-redundant S2648 set for a 5 fold cross-validation procedure is introduced to remove biases due to either the presence of the same protein or of the presence of the same protein position in both training and testing (Pires et al., 2014a). For sake of comparison, we also report the performances of the methods using a 5-fold cross validation made by random splitting the mutations in 5 sets ("random"). The "blind" test set consists of a subset of 351 mutations extracted from the original S2648 dataset, leaving the complement in the training set and comprising 2297 mutations.

The data set of P53 variations was also introduced by Pires et al. (2014a) as a case study, and consists of 42 variations within the DNA binding domain of the tumor suppressor protein p53, whose thermodynamic effects have previously been experimentally characterized and collected by several authors (see Pires et al., 2014a and references therein). When assessing the performances on this dataset, our method was trained on the subset of 2643 variations obtained by excluding 5 variations of P53 protein from the original S2648 dataset. This was done in order to remove the bias due to the presence of the same chain into the training set.

We also introduce the thermodynamic reversibility of the mutations, i.e. we consider that the inverse variation in a protein (e.g. GA and AG) is characterized by the negative value of the experimentally detected ΔΔG (Capriotti et al, 2008). By this, we both recast the thermodynamic property of the problem (ΔΔG(A,B)=- ΔΔG(B,A)) and balance the distribution of the available experimental measurements of free energy changes (Capriotti et al. 2008).

### 2.2 The INPS machine learning algorithm

INPS is based on a Support Vector Regression (SVR) as implemented by the libsvm package (Chang et al., 2011). In order to reduce the number of hyper-parameters of the SVR, we tested only the linear and the Radial Basis Function (RBF) kernels. In both cases, we used all the default parameters of the SVR, with the exception of C and γ. For the linear kernel, we optimized only the parameter C, which controls the trade-off between the margin width and the classification error on the training set. For the RBF kernel, both the values of C and γ (γ represents the inverse of the width of the RBF kernel, roughly defining the area of influence of a support vector, Chang et al., 2011) were tuned using a grid search procedure.

### 2.3 INPS: input encoding

INPS predictor consists of a SVR trained on the S2648 dataset using seven features of two kinds: 1) six descriptors encode the mutation type; 2) one descriptor encodes the evolutionary information. The variation of residue w with m is encoded with six real numbers:

- one input for the substitution w->m (BL62), scored with the Blosum62 matrix (Henikoff and Henikoff, 1992);
- two inputs for the hydrophobicity of the native (Hyw) and the mutant (Hym) residues rated with the Kyte-Doolitle scale (Kyte and Doolittle, 1982);
- one input to account for the mutability of the native residue (Mb) scored with the Dayhoff mutability scale (Dayhoff et al., 1978);

• two inputs (MWm, MWw) representing the molecular weights of the native and the mutant residues, respectively.

The evolutionary information is derived by analyzing the multiple sequence alignments of each query sequence obtained by running jackhmmer (Eddy, 2011) against the UNIREF90 dataset (release September, 2014). The parameters were set to: -N 3, -E 0.001, --domE 0.001, --incE 0.001, --incdomE 0.001. For each query protein, its multiple sequence alignment was processed to compute two different scores, derived from a sequence profile and a HMM model, respectively.

Both scores are separately adopted and tested. Specifically, concerning the "profile" score, the evolutionary information value is encoded by taking the difference between the wild-type (w) and the mutant (m) residues at the position of the profile where the mutation occurs. More formally, if P[k][a] represents the frequency of the residue a at the k-th position of the sequence profile, the "profile" score is computed as P[k][w]-P[k][m].

As an alternative (the so-called "HMM" score), the evolutionary information was encoded by means of a HMM model obtained by running the hmmbuild program from the HMMER suite (Eddy, 1998) on the multiple sequence alignment. Both the native and the mutated sequences are then aligned to the HMM with the hmmsearch program and the difference of the scores is taken as an estimation of the variation distance between the two sequences (HMM).

**Table 2.** Prediction performance on S2468 adopting a "per-protein (prot)" fold-cross-validation.

| Encoding* | Pearson Correlation | Standard Error (kcal/mol) |
|---|---|---|
| Mut | 0.41 | 1.32 |
| Mut+Profile | 0.50 | 1.28 |
| Mut+HMM | 0.52 | 1.26 |
| Mut+Profile+HMM | 0.51 | 1.27 |
| Mut+HMM-BL62 | 0.51 | 1.28 |
| Mut+HMM-Hy | 0.37 | 1.37 |
| Mut+HMM-Mb | 0.51 | 1.28 |
| Mut+HMM-MW | 0.50 | 1.28 |

*Mut=Bl62+Hy+Mb+Mw (see legend to Table 1). BL62=mutation scored according to the Blosum62 matrix. Hy=Hydrophobicity values for the wild-type and the mutant residues (Kyte and Doolittle, 1982). Mb=mutability value for the wild-type residue (Dayhoff et al., 1978). MW=molecular weight for the wild-type (MWw) and mutated (MWm) residues. Profile=difference between the sequence profile positions of the wild-type and mutated residues. HMM= HMM score of the wild-type protein and the mutated protein computed using HMMER program. (Eddy 1998). For (prot) definition see Materials and Methods (2.1).

## 3 RESULTS

### 3.1 Computing the change of protein stability upon residue substitution

For sake of clarity, we first evaluated to which extent each feature contributes information to the problem of computing changes in ΔG values upon residue substitution in the protein sequence (without using the inverse mutations). In Table 1, we list the Pearson correlation value of the various selected features with respect to the real-valued ΔΔGs in the S2648 set. All the selected features carry information different from random (the p-values associated to the correlation coefficients are significant). When encoding the variation type, the Blosum62 (BL62) substitution matrix score is the feature with the lowest correlation value, whereas the Hydrophobic difference between mutated and wild type residues (ΔHy =Hym-Hyw) appears the most relevant to infer the ΔG difference. Among the features encoding the evolutionary information, the HMM score is more informative than the profile score. We trained different SVRs (see Materials and Methods for details) and we adopted the more stringent 5-fold cross-validation split previously described (Pires et al., 2014a) to evaluate the method performance as a function of the different input features. The evaluation considered both the Pearson correlation and the standard error between the real and the predicted ΔΔG values.

Table 2 lists the correlation values obtained for each different input feature, after performing an optimization of the parameters on the training set with a grid search (without using the inverse mutations). In the first line (Mut) we report the correlation of the method when only the features encoding the variation type are included (namely, BL62+Hy+Mb+MW). In rows from two to four we added the features encoding the evolutionary information (Profile and HMM). When the evolutionary information is included, the Pearson correlation coefficient values increase by ten percentage points. The best predictor on the S2648 dataset is obtained with the inclusion of the HMM score (line three in Table 2). We also evaluated the performance of the method (Mut+HMM) by excluding step by step each of the components of the variation encoding (last four lines in Table 2). The performance only decreases when the hydrophobic information is not included in the input (the Pearson correlation and the standard error fall significantly, line six in Table 2). This finding corroborates the notion that the hydrophobic information is very relevant for the predictions of the ΔΔG values.

It is also worth mentioning that the performances of the method (for the different input encodings) are very stable for a wide range of SVR parameter values. Indeed, when the SVM parameters change, the Pearson correlation in cross-validation only ranges in the interval 0.50-0.52, with a corresponding standard error of 1.28-1.26 KCal/mol (see Supplementary Materials).

## 3.2 Inverse variations

From the thermodynamic point of view, an experimentally determined ΔΔG value should hold in a protein for a variant and its reverse (Capriotti et al., 2008): a protein and its variant should be endowed with the same free energy change, irrespectively of the reference protein (native or variant). If this is so, we can assume that the absolute value of free energy change is the same in going from one molecule to the other and that what changes is only the ΔΔG sign. By this, given a free energy value derived experimentally from a protein variation, we can take advantage of the previous statement and use the inverse variation (namely the variation that transforms back the variant into the original protein) by considering the value of the experimental measure with the opposite sign (-ΔΔG). Here we exploit this fact by testing the effect of adding the inverse variations to the cross-validation training and/or testing sets.

In Table 3, we show the results obtained by comparing the best performing method reported Table 2 (Mut+HMM) with its retrained version that includes the inverse variations in the learning sets. It appears that the method trained by including the inverse variations performs better in terms of correlation also when evaluated on test sets that do not contain them (Table 3, first column). When the inverse variations are included in the test sets (Table 3, last column) the training containing both direct and inverse variations performs significantly better as proven by the increase of both index values (Pearson correlation and mean square error). This indicates that the added (anti-symmetric) information can be helpful to stabilize the method and balancing it. For this reason, INPS is the predictor version trained using "both" direct and inverse variations. However, for comparing with other predictors, we always test the method only on the observed and experimentally detected free energy changes upon variations (direct values).

In Figure 1, we plot the cross-validation values of the 2648 experimentally detected free energy changes upon variations with respect to INPS predictions (INPS is trained on both direct and inverse ΔΔGs). It worth noticing that simply fitting the predicted versus the experimental ΔΔGs with a line crossing the origin, obtains a slope very close to 1 (Figure 1).

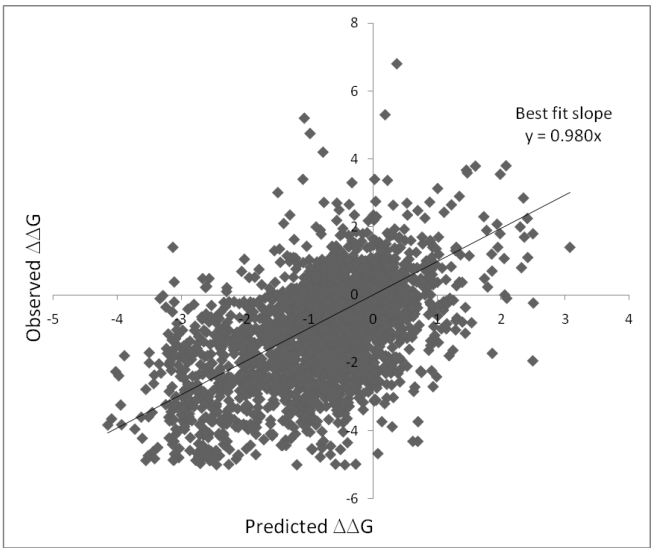


**Fig.1** Predicted versus observed free energy changes (ΔΔG) upon single point variations. The predictions are obtained using the per-protein five-fold cross-validation on S2648 (as described by Pires et al., 2014a).

## 3.3 Input alignment and performance

INPS relies on the multiple sequence alignment that is built to compute the HMM model to derive the scores of the wild-type and mutated sequences. This implies that we may aspect that the performances can be affected by the alignment quality and size. In Figure 2, we plot the graph of the Pearson correlation as a function of the number of sequences aligned to build the HMM model. The variations contained in the S2648 test set (as listed in the prot-folds by Pires et al., 2014a) are grouped according to the number of the aligned sequences relative to the protein whose variations are referring to. We group them in five different bins (Figure 2), then for each subset we computed the Pearson correlation between the predicted and the observed values. The figure shows that when the number of aligned proteins is larger than $10^5$, we may expect a performance higher than the average (0.53 in Table 3). On the contrary, when the number of aligned sequences falls below 100, the performance is lower than expected (Figure 2).

**Table 3**. Prediction performance on S2648 adopting "prot" fold-cross-validation and inverse mutations.

| S2468 Training | Testing Obs Mut Corr / SE | Testing Obs+Rev Mut Corr / SE |
|---|---|---|
| Obs Mut | 0.52 / 1.26 | 0.61 / 1.48 |
| Obs+Rev Mut | 0.53 / 1.29 | 0.69 / 1.29 |

Corr= Pearson correlation. SE= standard error (kcal/mole). Obs=S2648 observed experimental data. Obs+Rev= S2648 observed and inverse variations (5296 variations).

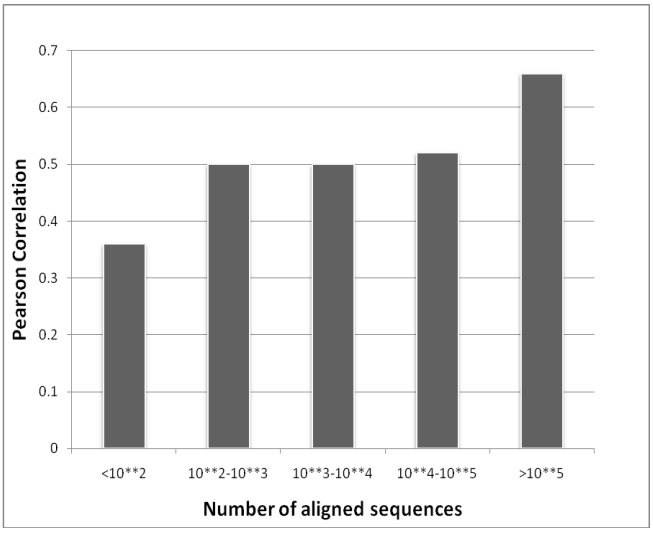


**Fig.2**. Predictive performance as a function of the number of aligned sequence in the corresponding protein multiple sequence alignment.

### 3.4 Comparison with existing methods

Recently three state-of-the-art and structure-based methods have been described, mCSM (Pires et al., 2014a), Duet (Pires et al., 2014b) and NeEMO (Giollo et al., 2014). Our purpose here is to compare INPS that is sequence based, with the best performing structure-based methods that have been routinely outperforming the sequence-based ones. For sake of comparison on the specific task of predicting ΔΔG values upon residue substitution, we focus on the same five splitting sets of S2468 that were previously described (Pires et al, 2014a) and are briefly introduced in the Material and Methods section. Apparently, when the split of the training set for the 5 fold cross validation procedure is done random (rand), per protein (prot) or per position (pos) (Pires et al, 2014a), the stringency of the training versus testing procedure changes. The per-protein split introduces less similarity in the training/testing set partition than the per-position split and even less when compared with the random one. Therefore, we also tested our INPS using the same partitions of the S2468 set, as done before (Pires et al., 2014a). In Table 4 we show that our sequence based INPS performs as well as the structure based mCSM adopting the same strategy for the stringent cross validation procedure (first and second column of Table 4) and well compares when the random split is adopted (third column in Table 4). On the blind set, also previously introduced (Pires et al., 2014a), INPS scores quite well when compared to the state-of-the-art predictors (mCSM, PopMusic and Duet). This is so even when all the methods are benchmarked on the P53 set (fifth column in Table 4). On this set, our sequence based method outperforms or performs similarly to all the-state-of the-art structure-based methods. IStable (Chen et al., 2013) which is a meta-predictor that exploits the predictions of several structure-based methods (I-Mutant2.0, AUTO-MUTE, MUPRO, PoPMuSiC2.0, CUPSAT) is not able to achieve the single-method performances of mCSM and INPS. For the recently introduced Duet (to combine two structure-based methods: SDM and mCSM, see Pires at al., 2014b), the mean standard error reduces to 1.39 kcal/mol, while correlation is still high. However, SDM and mCSM are both structure-based methods and do not include evolutionary information. Then, we simply combined INPS with mCSM or Duet by averaging the ΔΔG values provided by the two tools. In these cases, the prediction level further improves and achieves the highest performance obtained on the P53 dataset (last two lines in Table 4). The result indicates that INPS carries a complementary and useful information with respect to the structure-based methods.

Adopting the only ΔΔG sign, INPS predictions can be interpreted to identify stabilizing and destabilizing variations. Although this is not the optimal solution to develop a classifier of stability variations, in this task INPS performs equally well or better than the most recent methods on the P53 dataset (See Table 1S, supplementary materials). In particular when the Matthews correlation coefficient is taken into account INPS outperforms the state-of-the-art methods (when evaluated on the P53 set, Table 1S, supplementary materials). Given the paucity of the available data to benchmark the different methods using "real" blind sets (in this case only 42 variations), the results reported in our paper should be considered only indicative of how the different methods score.

In Figure 3, we plot the predicted versus the observed values of ΔΔG for the P53 dataset (dark diamonds). In order to highlight the ability to predict anti-symmetrically the inverse variations we also plot the anti-symmetric pairs in gray squares (Figure 3). It is evident that INPS learned the thermodynamic anti-symmetric property. The property is equivalent to a rotation of 180° around the origin and after the rotation, the dark diamonds superimpose quite well the gray squares.

In case we add the predictions of the inverse mutation the Pearson correlation and the mean square errors become 0.80 and 1.51, respectively.

## 4 DISCUSSION AND CONCLUSIONS

In this paper, we introduce INPS, a new method only based on sequence information, for predicting the effect of variations on protein stability. The novelty is that the method takes the protein sequence as input and reaches a performance that is similar to that achieved by adopting protein structures. Therefore, our INPS can be used in the large amount of cases in which protein structure is not available. We showe that thanks to the evolutionary information, in the form of HMM, INPS performance is similar to those of the best-performing structure-based methods. Furthermore, we showe that INPS predictions are complementary to those generated by the best-performing mCSM and Duet structure-based predictors (at least when tested on the new p53 set). When mCSM or Duet are combined with INPS, by averaging their ΔΔG values, the overall performance significantly improves. However and most importantly, we show that with INPS and starting from the protein sequence, it is possible to obtain a reliable prediction for inferring the effect of variations on protein stability. This can be useful when annotating protein variants when the protein structure is not available and/or after detecting non-synonymous single nucleotide polymorphisms during massive sequencing experiments.

**Table 4**. Comparison with state-of-the-art methods on different datasets

| Method | S2468 5-fold-prot Corr / SE | S2468 5-fold-pos Corr / SE | S2468 5-fold-rand Corr / SE | Blind-set Corr / SE | P53 set Corr / SE |
|---|---|---|---|---|---|
| INPS | 0.53 / 1.29 | 0.54 / 1.28 | 0.60 / 1.22 | 0.68 / 1.26 | 0.71 / 1.49 |
| SDM * | – / -- | – / -- | | – / -- | 0.29 / 1.75 |
| mCSM* | 0.51 / 1.26 | 0.54 / 1.23 | 0.69 / 1.05 | 0.67 / 1.19 | 0.68 / 1.40 |
| PopMusic2.0* | – / -- | – / -- | 0.63 /1.15 | 0.73 / 1.09 | 0.56 / 1.52 |
| IStable* | – / -- | – / -- | | – / -- | 0.49 / 1.59 |
| Duet^ | – / -- | – / -- | | 0.71 / 1.13 | 0.68 / 1.39 |
| NeEMO | – / -- | – / -- | – / -- | – / -- | 0.47 / 1.65 |
| INPS+mCSM | – / -- | – / -- | | – / -- | 0.75 / 1.39 |
| INPS+Duet | – / -- | – / -- | | -- /-- | 0.75 / 1.35 |

Corr= Pearson correlation. SE= standard error (kcal/mole). * Data are taken from Pires et al., 2014a and ^ from Pires et al., 2014b. 5-fold cross-validation of S2648 is tested using 3 different split: “prot” is per protein split, “pos” is per position split, “random” is random split of the mutations as described in Material and Method section.

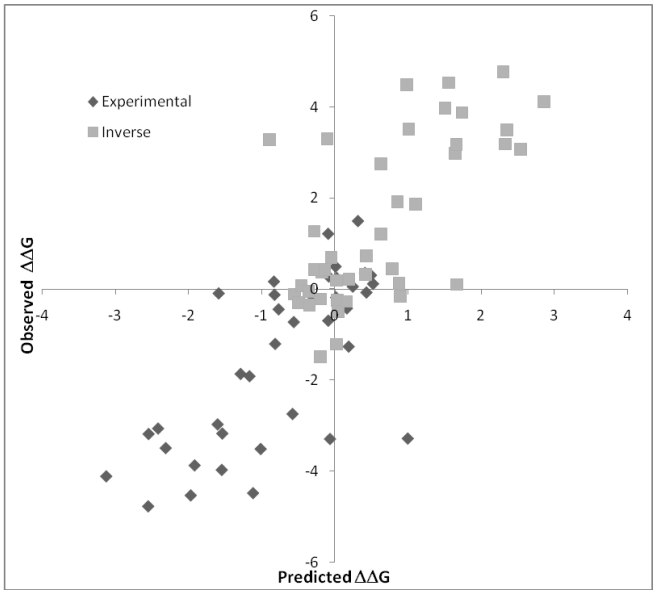


**Fig.3**. Predicted versus observed free energy variations upon single point variations of P53.

## ACKNOWLEDGEMENTS

*Funding*: PRIN 2010-2011 project 20108XYHJS (to P.L.M.) (Italian MIUR); COST BMBS Action TD1101 and Action BM1405 (European Union RTD Framework Program, to R.C); PON projects PON01_02249 and PAN Lab PONa3_00166 (Italian Miur to R.C. and P.L.M.); FARB-UNIBO 2012 (to R.C.).

## REFERENCES

Capriotti, E., et al. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server issue), 306–310.
Capriotti, E. et al. (2008) A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics, 26;9 Suppl 2:S6.
Capriotti,E., et al. (2012) Bioinformatics for personal genome interpretation. Brief Bioinform., 13, 495–512.
Casadio,R., et al. (2011) Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. Hum. Mutat., 2, 1161–1170.
Chang,C.C. et al. (2011) LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27.
Chen,C.W et al., (2013) iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics 14 Suppl 2:S5.
Cheng, J., et al. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62, 1125–1132.
Dehouck,Y. et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics, 25, 2537–2543.
Dayhoff,M.O. et al. (1978) A model of evolutionary change in proteins. In "Atlas of Protein Sequence and Structure, vol. 5, suppl. 3.
Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics. 14:755-763.
Eddy, S.R. (2011) Accelerated Profile HMM searches. PLOS Comp Biol. 7:e1002195
Giollo, M. et al. (2014) NeEMO: A Method Using Residue Interaction Networks to Improve Prediction of Protein Stability upon Mutation. BMC Genomics, 15(Suppl 4):S7
Guerois,R. et al. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol., 320, 369–387.
Henikoff,S. and Henikoff,J.G. (1992) Amino Acid Substitution Matrices from Protein Blocks. PNAS 89: 10915–10919.
Huang, L.T., et al. (2007) iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics 23, 1292–1293.
Hudson,T.J., et al. (2012) International network of cancer genome projects. Nature, 464, 993–998.
Khan, S., Vihinen, M. (2010) Performance of protein stability predictors. Hum Mutat 31(6), 675–684.
Kumar,M.S. et al. (2006) Protherm and pronit: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res., 34 (Suppl. 1), D204–D206.
Kyte,J., and Doolittle R.F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105-132.
Lahti,J.L., et al., T (2012) Bioinformatics and variability in drug response: a protein structural perspective. J. R. Soc. Interface, 9, 1409–1437.
Masso,M. and Vaisman,I. (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics, 24, 2002–2009.

Parthiban,V., et al. (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34 (Web Server issue), 239–242.
Pires,D.E.V., et al. (2014a) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3), 335–342.
Pires,D.E.V., et al. (2014b) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42(Web Server issue):W314-319.
Rahman, N.A. (1968) A Course in Theoretical Statistics, Charles Griffin and Company, 1968.
Stratton,M.R., et al. (2012) The cancer genome. Nature, 458, 719–724.
Topham,C.M. et al. (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng., 10, 7–21.
Teng,S., et al. (2010) Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics 11 Suppl 2, 5.
Wainreb,G (2011) Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. Bioinformatics 27,3286-3292.
Worth,C.L. et al. (2011) SDM – a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res., 39 (Suppl. 2), W215–W222.
Yin,S., er al. (2007) Eris: an automated estimator of protein stability. Nat Methods 4, 466–467.
Zhou, H., Zhou, Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11, 2714–2726.

# Supplementary Materials for: "INPS: Predicting the Impact of Non-Synonymous Variations on Protein Stability from Sequence"

Piero Fariselli, Pier Luigi Martelli, Castrense Savojardo, Rita Casadio

## INPS used as a classifier

All the available predictorsof protein stability changes upon single point variations have been trained using subsets of the Protherm data (Kumar et al., 2006). It is then very difficult to benchmark the different predictors using an independent dataset. To partially overcome this problem, Pires et al., 2014a, evaluated the methods on a set consisting of 42 variations within the DNA binding domain of the tumor suppressor protein p53, whose thermodynamic effects have previously been experimentally characterized (variations not present in theProtherm database).
For sake of comparison, we benchmarked INPS with the most recent and available predictors on the task of discriminating between destabilizing and stabilizing P53 mutations. INPS has been trained without using any of the P53 variations (nor other variations in the same proteins) with a Support Vector Regression. To evaluate INPS as a classifier, we used the sign of INPS $\Delta\Delta G$ predictions as class label (+1 for stabilizing, -1 for destabilizing variations). In table 1S, we list the performances as compared with those of the most recent methods.

**Table 1S**. Comparison on two-state prediction with different methods on the P53 set

| Method | MCC | Q | P(+) | S(+) | P(-) | S(-) |
|---|---|---|---|---|---|---|
| **INPS** | 0.46 | 0.76 | 0.53 | 0.72 | 0.88 | 0.77 |
| **MuProSVM** | 0.07 | 0.69 | 0.33 | 0.18 | 0.75 | 0.87 |
| **Imutant3** | 0.05 | 0.67 | 0.2 | 0.1 | 0.73 | 0.87 |
| **AutomuteRF** | 0 | 0.74 | 0 | 0 | 0.74 | 1 |
| **isStable** | 0 | 0.26 | 0.26 | 1 | 0 | 0 |
| **Duet** | 0.36 | 0.79 | 0.75 | 0.27 | 0.79 | 0.97 |
| **mCSM*** | 0.38 | 0.79 | 1 | 0.18 | 0.78 | 1 |
| **SDM*** | 0.07 | 0.62 | 0.31 | 0.36 | 0.76 | 0.71 |
| **PopMusic2*** | 0.12 | 0.69 | 0.38 | 0.27 | 0.76 | 0.84 |
| **NeEMO** | 0.09 | 0.64 | 0.17 | 0.09 | 0.72 | 0.84 |

INPS, Duet (Pires et al., 2014b), mCSM (Pires et al., 2014a), SDM (Worth et al., 2011), PopMusic2 (Dehouck et al., 2009) and NeEMO (Giollo et al. 2014) the classes are assigned according to the sign of the predicted $\Delta\Delta G$ values.
MuProSVM (Cheng et al., 2006), Imutant3 (Capriotti et al., 2008), AutomuteRF (Masso and Vaisman 2008) and iStable (Chen at al. 2013) are evaluated using the corresponding web server class predictions. * Data are taken from Pires et al., 2014a

For comparing with class-based predictors, here we introduce others performance measures. Considering the positive and negative classes on the bases of the sign of the $\Delta\Delta G$ values (predicted and observed), performances are evaluated adopting the following scoring indexes:
The overall accuracy is:

$$Q = \frac{C}{T}$$

Where $C$ is the total number of correctly predicted variations and $T$ is the total number of variations. The Matthew's correlation coefficient $MCC$is defined as:

$$MCC(s) = \frac{p(s)n(s) - u(s)o(s)}{W}$$

where $W$ is the normalization factor:

$$W = \sqrt{[(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))]}$$

for each class $s$ (stabilizing or destabilizing mutations); $p(s)$ and $n(s)$ are the total number of correct predictions and correctly rejected assignments, respectively. $u(s)$and $o(s)$ are the numbers of false negatives and false positives for the $s$ class.
The coverage $S$ (sensitivity) for each discriminated class $s$is evaluated as:

$$S(s) = \frac{p(s)}{p(s) + u(s)}$$

The probability of correct predictions $P$ (or positive predictive values or specificity) is computed as:

$$P(s) = \frac{p(s)}{p(s) + o(s)}$$

## References


Capriotti, E. et al. (2008) A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics, 26;9Suppl 2:S6.

Chen,C.W et al., (2013) iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics 14 Suppl 2:S5.

Cheng, J., et al. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62, 1125–1132.

Dehouck,Y. et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics, 25, 2537–2543.

Giollo, M. et al. (2014) NeEMO: A Method Using Residue Interaction Networks to Improve Prediction of Protein Stability upon Mutation. BMC Genomics, 15(Suppl 4):S7

Kumar,M.S. et al. (2006) Protherm and pronit: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res., 34 (Suppl. 1), D204–D206.

Masso,M. and Vaisman,I. (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics, 24, 2002–2009.

Pires,D.E.V., et al. (2014a) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3), 335–342.

Pires,D.E.V., et al. (2014b) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42(Web Server issue):W314-319.

Worth,C.L. et al. (2011) SDM – a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res., 39 (Suppl. 2), W215–W222.