Structural bioinformatics

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

Piero Fariselli\(^1\), Pier Luigi Martelli\(^1\), Castrense Savojardo\(^1\) and Rita Casadio\(^1\)

\(^1\)Biocomputing Group, Department of Biology, University of Bologna, 40126 Bologna and \(^2\)Department of Computer Science and Engineering, University of Bologna, 40127 Bologna, Italy

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on February 20, 2015; revised on April 15, 2015; accepted on May 2, 2015

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 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 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-the-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 and implementation:** The presented method is available as web server at http://inps.biocomp.unibo.it.

**Contact:** piero.fariselli@unibo.it

**Supplementary information:** Supplementary Materials are available at *Bioinformatics* online.

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
The available methods were trained under different conditions and on different datasets. 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 ProTherm 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 article, 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 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 et al., 2014b).

2 Materials and methods

2.1 Datasets

In this article, 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 dataset comprises 2648 single-point variations in 132 different globular proteins. In this article, 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 are 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, comprising 2297 mutations.

The dataset 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 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). 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: (i) six descriptors encode the mutation type; (ii) 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-Doolittle 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 (MWW, MWM) 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 jackhammer (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 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 hmmbuild program and the difference of the scores is taken as an estimation of the variation distance between the two sequences (HMM).

### 3 Results

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

For sake of clarity, we first evaluated to what extent each feature contributes information to the problem of computing changes in $\Delta G$ 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 $\Delta G$s 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 ($\Delta H = H_{m} - H_{w}$) appears the most relevant to infer the $\Delta G$ difference. Among the features encoding the evolutionary information, the HMM score is more informative than the profile score. We trained different SVRs (see Section 2 for details) and we adopted the most stringent 5-fold cross-validation split previously described (Pires et al., 2014) 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 $\Delta G$ 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 + $H_{m}$ + $m$ + 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 S2468 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 $\Delta G$ 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 $\Delta A G$ 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 absolute free energy change, irrespective 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 $\Delta G$ sign. By this, given a free energy value derived experimentally

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pearson correlation</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL62</td>
<td>0.11</td>
<td>1E–8</td>
</tr>
<tr>
<td>$\Delta H$</td>
<td>0.28</td>
<td>6.0E–49</td>
</tr>
<tr>
<td>$M_B$</td>
<td>0.17</td>
<td>3.0E–19</td>
</tr>
<tr>
<td>$\Delta$W/WW</td>
<td>0.18</td>
<td>3.0E–21</td>
</tr>
<tr>
<td>Profile</td>
<td>0.21</td>
<td>1.0E–24</td>
</tr>
<tr>
<td>HMM</td>
<td>0.27</td>
<td>5.0E–44</td>
</tr>
</tbody>
</table>

$w$, 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); $\Delta H$ (Hym-Hyw), hydrophobicity difference between mutated and wild type residues (Kyte and Doolittle, 1982); $M_B$, mutability value for the wild-type residue (Dayhoff et al., 1978); $\Delta W$ (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.

<table>
<thead>
<tr>
<th>Encoding*</th>
<th>Pearson correlation</th>
<th>Standard error (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mut</td>
<td>0.41</td>
<td>1.32</td>
</tr>
<tr>
<td>Mut + Profile</td>
<td>0.50</td>
<td>1.28</td>
</tr>
<tr>
<td>Mut + HMM</td>
<td>0.52</td>
<td>1.26</td>
</tr>
<tr>
<td>Mut + Profile + HMM</td>
<td>0.51</td>
<td>1.27</td>
</tr>
<tr>
<td>Mut + HMM-BL62</td>
<td>0.51</td>
<td>1.28</td>
</tr>
<tr>
<td>Mut + HMM-Hy</td>
<td>0.37</td>
<td>1.37</td>
</tr>
<tr>
<td>Mut + HMM-Mb</td>
<td>0.51</td>
<td>1.28</td>
</tr>
<tr>
<td>Mut + HMM-MW</td>
<td>0.50</td>
<td>1.28</td>
</tr>
</tbody>
</table>

*Mut, BL62 + $H_{m}$ + $M_B$ + $M_W$ (see legend to Table 1); BL62, mutation scored according to the Blosum62 matrix; $H_{m}$, Hydrophobicity values for the wild-type and the mutant residues (Kyte and Doolittle, 1982); $M_B$, mutability value for the wild-type residue (Dayhoff et al., 1978); $M_W$, molecular weight for the wild-type (WW) and mutated (WWm) 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 Section 2.1.
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 ($/C_0\Delta G$). 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.

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.

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 $\Delta G$s). It worth noticing that simply fitting the predicted versus the experimental $\Delta G$s with a line crossing the origin, obtains a slope very close to 1 (Fig. 1).

### 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 (Fig. 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 10⁵, the performance is 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 (Fig. 2).

### 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 $\Delta G$ 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 Section 2. 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
Table 4. Comparison with state-of-the-art methods on different datasets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INPS</td>
<td>0.53/1.29</td>
<td>0.54/1.28</td>
<td>0.60/1.22</td>
<td>0.68/1.26</td>
<td>0.71/1.49</td>
</tr>
<tr>
<td>SDM</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>0.29/1.75</td>
</tr>
<tr>
<td>mCSM</td>
<td>0.51/1.26</td>
<td>0.54/1.23</td>
<td>0.69/1.05</td>
<td>–/–</td>
<td>0.68/1.40</td>
</tr>
<tr>
<td>PopMusic2.0</td>
<td>–/–</td>
<td>–/–</td>
<td>0.63/1.15</td>
<td>0.73/1.09</td>
<td>0.56/1.52</td>
</tr>
<tr>
<td>IStable</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>0.49/1.59</td>
</tr>
<tr>
<td>Duet</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>0.71/1.13</td>
<td>0.68/1.39</td>
</tr>
<tr>
<td>NeEMO</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>0.47/1.65</td>
</tr>
<tr>
<td>INPS + mCSM</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>0.75/1.39</td>
</tr>
<tr>
<td>INPS + Duet</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>0.75/1.35</td>
</tr>
</tbody>
</table>

Corr, Pearson correlation; SE, standard error [kcal/mole].

aData are taken from Pires et al. (2014a) and b from Pires et al. (2014b). 5-fold cross-validation of S2648 is tested using three different splits: ‘prot’ is per protein split, ‘pos’ is per position split, ‘random’ is random split of the mutations as described in Material and Method section.

The result indicates that INPS carries a complementary and useful information with respect to the structure-based methods.

Using the 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).

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 large amount of cases in which protein structure is not available. We showed 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 showed 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 to annotate protein variants when the protein structure is not available and/or after detecting non-synonymous single nucleotide polymorphisms during massive sequencing experiments.

4 Discussion and conclusions

In this article, 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 large amount of cases in which protein structure is not available. We showed 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 showed 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 to annotate protein variants when the protein structure is not available and/or after detecting non-synonymous single nucleotide polymorphisms during massive sequencing experiments.
Funding
PRIN 20102011 project 20108XHYHS (to P.L.M.) (Italian MIUR); COST BMBS Action TD1101 and Action BMI405 (European Union RTD Framework Program, to R.C); PON projects PON01_02249 and PAN Lab PONa3_00166 (Italian Muir to R.C. and P.L.M.); FARB-UNIBO 2012 (to R.C.).

Conflict of Interest: none declared.

References