Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0

Yves Dehouck1,∗, Aline Grosfils2, Benjamin Folch1, Dimitri Gilis1, Philippe Bogats2 and Marianne Rooman1
1Bioinformatique génomique et structurale and 2Modélisation et contrôle de bioprocédés, Université Libre de Bruxelles, Av Fr. Roosevelt 50, CP165/61, 1050 Brussels, Belgium

Received on March 18, 2009; revised on July 7, 2009; accepted on July 15, 2009
Advance Access publication August 3, 2009

1 INTRODUCTION

The ability to design more stable and robust proteins, that maintain the stability of the change in folding free energy upon mutation leads to the protein’s structure and function. The experimental determination of such stability too much, for they may otherwise cause the loss of the structure, stability and function of proteins, such as for example their solubility or their specificity to tune numerous other physicochemical or biological properties. It can also be interesting in silico methods able to predict rapidly, and with reasonable accuracy, the stability changes resulting from all possible mutations in a protein.

Results: We exploit newly developed statistical potentials, based on a formalism that highlights the coupling between four protein sequence and structure descriptors, and take into account the amino acid volume variation upon mutation. The stability change is expressed as a linear combination of these energy functions, whose proportionality coefficients vary with the solvent accessibility of the mutated residue and are identified with the help of a neural network. A correlation coefficient of \( R = 0.68 \) and a root mean square error of \( \sigma_c = 1.15 \text{ kcal/mol} \) between measured and predicted stability changes are obtained upon cross-validation. These scores reach \( R = 0.79, \) and \( \sigma_c = 0.86 \text{ kcal/mol} \) after exclusion of 10% outliers. The predictive power of our method is shown to be significantly higher than that of other programs described in the literature.

Availability: http://babylone.ulb.ac.be/popmusic
Contact: ydehouck@ulb.ac.be
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The ability to design more stable and robust proteins, that maintain their activity under unusual pH or temperature conditions, would be a precious asset in many industrial sectors. It can also be interesting to tune numerous other physicochemical or biological properties of proteins, such as for example their solubility or their specificity to various ligands. The considered mutations should not alter the stability too much, for they may otherwise cause the loss of the protein’s structure and function. The experimental determination of the change in folding free energy upon mutation leads to the most reliable information, but is costly and time consuming, and therefore not adapted to explore extensively the mutational space of a protein. Predictive methods, capable of estimating rapidly the stability changes resulting from all possible mutations, are thus to become unavoidable in the implementation of any protein engineering project. The development of such methods also bears a considerable interest at the fundamental level, as it participates to a better understanding of the relationships between the sequence, structure, stability and function of proteins.

The earliest theoretical approaches for predicting stability changes caused by mutations were focused on a small number of mutations in a single protein, usually even at a single site (e.g. Basch et al., 1987; Eriksson et al., 1992; Miyazawa and Jernigan, 1994; Tidor and Karplus, 1991; van Gunsteren and Mark, 1992). The PoPMuSiC program (Gillis and Rooman, 2000; Kwasigroch et al., 2002) was the first to be general enough for predicting stability changes caused by mutations at any point in any protein. It relies on knowledge-based potentials, based on distances between residue pairs or backbone torsion angles. Since then, several other predictive methods have been developed. Some are also based on statistical potentials (Parthiban et al., 2006; Zhou and Zhou, 2002), while others rely on empirical potentials describing physically the interactions that contribute to protein stability (Guerois et al., 2002; Yin et al., 2007).

Different types of machine learning tools, mostly support vector machines, have recently also been used in this context (Capriotti et al., 2005; Cheng et al., 2006; Masso and Vaisman, 2008). Note that all the above-mentioned methods require the structure of the target protein. Machine learning tools were also used to develop methods predicting stability changes in proteins of unknown structure, with reasonable but obviously more limited performances (Capriotti et al., 2005; Cheng et al., 2006; Masso and Vaisman, 2008).
Here, we keep the same basic idea, but design a completely new energy function. We use a whole new set of 24 statistical potentials, as well as terms modeling the volume changes upon mutation, and express the folding free energy change as a single linear combination of these terms, with weighting coefficients that depend on the solvent accessibility. To reproduce the smooth transition between the core and the surface of proteins, and to generalize the step functions used in the original version, we assume that these weighting coefficients have a sigmoid shape. The parameters defining these sigmoid functions are identified with the help of a neural network, which is also an important novelty of our algorithm, PoPMuSiC-2.0.

2 METHODS

2.1 Database of experimentally characterized mutations

We extracted, from the ProTherm database (Battistit et al., 2004), a set of mutations whose impact on the stability of the protein structure has been measured experimentally. These data were carefully checked by consulting the original articles and a number of erroneous inputs had to be corrected or removed. We considered only single-site mutations, in globular proteins whose experimental structure (either X-ray or NMR) is available. We excluded mutations introduced in pseudo wild-type constructs, as well as mutations in heme-proteins (except if the measurements were performed on the apo form of the protein, and the structure of this apo form is not affected by the mutations). Seven discrete domains of the mutation spectrum were considered, to be separated from another amino acid (whatever its nature or position). Mutations for which this information could not be retrieved from the original papers were eliminated. All remaining values were adapted to correspond to the folding free energy change per mole of monomer.

With these criteria, 2648 different point mutations, in 131 proteins, were selected. For some mutants, several measures of the variation in folding free energy (ΔΔG) have been performed, sometimes in different conditions. To avoid any redundancy in our database, we considered only one value per mutant, which is noted ΔΔGcrip and is defined as an average of all available measured ΔΔG values. Measurements performed with a pH close to 7, a temperature close to 25°C, and without additives, are given a higher weight in the averaging procedure. The complete database and details on this averaging procedure are given as Supplementary Material.

2.2 Simplified protein representation

The sequence of a protein is described by the nature of the amino acid at each position, i, noted ai. Its three-dimensional structure is represented by several descriptors. First, the conformation of a residue at position i is defined by its backbone torsion angles, ti. Seven discrete domains of ti are considered, grouping specific local organizations of the protein chain (Roosan et al., 1991). Secondly, the spatial distance between the two amino acids at positions i and j, noted dij, is referred to as dj. The distance values are computed between the average geometric centers of the side chains and distributed in bins of 0.2 Å width, from 3.0 to 8.0 Å (Dehouck et al., 2006). A last bin is added for dij’s below 3.0 Å. When dij exceeds 8.0 Å, we consider that the two residues do not interact, and no energy is computed. Finally, the solvent-accessibility of the amino acid at position ai is defined as the ratio of its solvent-accessible surface in the considered structure, as computed by DSSP (Kabsch and Sander, 1983), and in an extended tripeptide Gly-X-Gly (Rose et al., 1985). Five discrete values of ai are considered: 0-5, 5-15, 15-30, 30-50 and 50-100%.

Table 1. Selection of statistical potentials

<table>
<thead>
<tr>
<th>Group of potentials</th>
<th>Potentials included in the group</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1: Basic potentials</td>
<td>Local: $\Delta W_{t1}$, $\Delta W_{t2}$, $\Delta W_{dij}$, $\Delta W_{as}$, $\Delta W_{ci}$, $\Delta W_{cj}$, $\Delta W_{res}$, $\Delta W_{side}$, $\Delta W_{sol}$, $\Delta W_{surf}$, $\Delta W_{total}$</td>
</tr>
<tr>
<td>G2: Low-order coupling terms</td>
<td>Local: $\Delta W_{t1}$, $\Delta W_{t2}$, $\Delta W_{dij}$, $\Delta W_{as}$, $\Delta W_{ci}$, $\Delta W_{cj}$, $\Delta W_{res}$, $\Delta W_{side}$, $\Delta W_{sol}$, $\Delta W_{surf}$, $\Delta W_{total}$</td>
</tr>
<tr>
<td>G3: High-order coupling terms</td>
<td>Local: $\Delta W_{t1}$, $\Delta W_{t2}$, $\Delta W_{dij}$, $\Delta W_{as}$, $\Delta W_{ci}$, $\Delta W_{cj}$, $\Delta W_{res}$, $\Delta W_{side}$, $\Delta W_{sol}$, $\Delta W_{surf}$, $\Delta W_{total}$</td>
</tr>
<tr>
<td>G4: Volume terms</td>
<td>$\Delta F_{sol}$, $\Delta V_{surf}$</td>
</tr>
<tr>
<td>G5: Independent term</td>
<td>1</td>
</tr>
</tbody>
</table>

2.3 Estimation of the folding free energy changes

In order to describe the different interactions contributing to protein stability, we rely on a set of statistical potentials, extracted from a database of known protein structures. A form commonly used for such potentials is

$$\Delta W_{cr} = -kT \log \frac{P(c_1, c_2)}{P(c_1) P(c_2)}$$

where $c_1$ and $c_2$ are sequence or structure descriptors (i.e., $t_i$, $d_{ij}$, or $a_{ij}$) of the same amino acids or of neighboring ones, and P are their relative frequencies of occurrence in a large, non-redundant, dataset of protein structures. Some of us previously generalized this relationship to derive complex potentials describing the correlations between more than two descriptors, while ensuring that each contribution is counted only once (Dehouck et al., 2006):

$$\Delta W_{cr} = -kT \log \prod_{i=1}^{n} \frac{P(c_{i1}, c_{i2}, ..., c_{in})}{P(c_{i1}) P(c_{i2}) ... P(c_{in})}$$

where n is the number of descriptors. In this work, we use 24 different potentials, with n ranging from 2 to 7; they are listed in Table 1, and grouped in several subsets $G_i$ according to their complexity. They can be divided into two major classes: local and non-local potentials, which describe the correlations between descriptors attached to residues close to each other along the sequence, or close to each other in space, respectively. For example, $\Delta W_{as} \equiv \Delta W(t_1, t_2)$ reflects the influence of the nature of an amino acid $a_i$ on the conformation $t_j$ of a neighboring residue, while $\Delta W_{di} \equiv \Delta W(t_1, t_j, d_{ij})$ reflects the propensity of an amino acid $a_i$ in a conformation $t_j$, to be separated from another amino acid (whatever its nature and conformation) by a spatial distance $d_{ij}$. The main difference between these potentials and those used by other prediction algorithms is that they take into account simultaneously the correlations between several sequence and structure descriptors, offering thus a more complete picture of the complex ensemble of interactions that rule protein stability.

Another parameter used to predict the mutant stability is the volume difference $\Delta V$ between the mutant and wild-type amino acids. If the mutant amino acid is smaller ($\Delta V < 0$), a cavity is created, which usually destabilizes the protein (Eriksson et al., 1992). On the other hand, accommodating a large side-chain ($\Delta V > 0$) may induce stresses in the structure, which are also likely to have a destabilizing impact. Statistical potentials cannot describe correctly such effects, since they are derived from a dataset of native structures of wild-type proteins, with very few packing defects. As the amplitudes of
The inflection points
To identify the 27 neuron and a linear output (Fig. 1). Its mathematical expression is:

We use a 5-fold validation procedure: the neural network is trained on 4/5 of the mutant dataset and validated on the remaining 1/5. Five different runs are performed, so that each mutant is included in one of the validation sets. The performances are assessed using the root mean square error \( \sigma \) (Equation (6)) and the Pearson correlation coefficient \( R \) between the measured \( \Delta \Delta G_{\text{P}} \) and predicted \( \Delta \Delta G_{\text{P}} \) values of the folding free energy changes. We also performed a 10-fold validation and observed no significant differences, which indicates that our dataset is large enough for 5-fold validation.

\[ F_{0} = \sum_{i=1}^{27} \left( \frac{\Delta \Delta G_{\text{P}}}{\partial \theta_{i}} \right) \left( \frac{\Delta \Delta G_{\text{P}}}{\partial \theta_{i}} \right) \]

with \( \theta = \{ \theta_{A}, \theta_{B} \} \). The reduction procedure works as follows:

(1) After full identification of the model, the matrix \( F \) is computed.

(2) The parameter \( \theta_{i} \) with the highest variance \( E_{\theta_{0}} \) is identified, and the set of parameters \( \theta_{j} \) whose variance \( E_{\theta_{j}} \) is larger than 80% of this maximal value \( E_{\theta_{0}} \) is selected. This set represents the parameters that are least precisely identified in the model.

(3) Among this set of parameters, the parameter \( \theta_{i} \) which has the largest covariance \( E_{\theta_{i}} \) with another parameter \( \theta_{j} \) is eliminated. Indeed, a large covariance indicates that the parameters are redundant, and thus that the one with the large variance can safely be eliminated.

(4) When eliminated, hardly assessable weights \( (\nu_{i}) \) and biases \( (b_{i}) \) are cancelled out, but slopes and inflection points are set to their initial values \( (x=1 \text{ and } C=50) \). Note that the suppression of a weight \( \nu_{i} \) automatically implies the elimination of the corresponding \( \theta_{i} \) and \( r_{i} \).

(5) The model is then re-identified with the remaining parameters according to the procedure described above.

The reduction procedure is then iteratively repeated to eliminate other parameters. It is ended when none of the remaining parameters has a variance larger than 10 mol\(^2\)/kcal\(^2\) and a covariance with another parameter larger than 1 mol\(^2\)/kcal\(^2\); these values were selected by trial and error.

3 RESULTS

3.1 Training and validation

The neural network is trained on 4/5 of the mutant dataset and validated on the remaining 1/5. Five different runs are performed, so that each mutant is included in one of the validation sets. The performances are assessed using the root mean square error \( \sigma \) (Equation (6)) and the Pearson correlation coefficient \( R \) between the measured \( \Delta \Delta G_{\text{P}} \) and predicted \( \Delta \Delta G_{\text{P}} \) values of the folding free energy changes. We also performed a 10-fold validation and observed no significant differences, which indicates that our dataset is large enough for 5-fold validation.

When all potentials are considered, the total number of parameters is quite large (\( N_{\text{P}} \approx 85 \) after reduction). Still, the correlation coefficients and root mean square errors obtained on the training and validation sets are very similar (Table 2), indicating that the model is not over fitted. However, the results in Table 2 show that the potentials belonging to the group \( G_{2} \) (Table 1) are not necessary and that removing them does not decrease the performances in cross-validation. This is not surprising since these are high-order coupling terms, reflecting weak correlations that are not already accounted for by the lower-order couplings. In addition, these \( G_{2} \) potentials were previously shown to be very sensitive to the size of the database, and to be useful only in some specific applications (Dehouck et al., 2006). In contrast, the removal of either the low-order coupling terms (\( G_{2} \)), the volume terms (\( G_{3} \)), or the independent term (\( G_{4} \)) leads to...
We chose here to exclude, one by one, the 10% mutants that impact reaches so as to obtain comparable results. As can be seen in the last row stability changes as high as 0.78, and a root mean square error $\sigma$ dataset the mutants with a larger than 1.5 kcal/mol in each of the five runs performed with the proline, since they are likely to induce significant structural imperfections in our model, or to structural modifications provoked upon cross-validation. This can of course be attributed to a decrease in the predictive power. In summary, the best results are obtained by combining the basic potentials $G_1$, the low-order coupling terms $G_2$, the volume terms $G_4$, and the independent term $G_5$. This yields an $R_c$ value of 0.62 and a $\sigma_c$ value of 1.16 kcal/mol upon cross-validation.

Some mutations are consistently badly predicted, independently of the potentials chosen. This can of course be attributed to imperfections in our model, or to structural modifications provoked by the mutation and not taken into account by PoPMuSiC. But this can also be due to an experimental measure made in specific, non-physiological, conditions or affected by a significant error, to a poorly resolved structure, or to mistakes in the database indexing of the measured $\Delta G$ value. To account for this, we also computed the performance measures after removal of these outliers, with the aim of giving a more relevant evaluation of the predictive power. We chose here to exclude, one by one, the 10% mutants that impact most negatively the correlation coefficient. After this exclusion, we obtain a correlation coefficient $R_c$ between predicted and measured $\Delta G_0$, in direct ($R_d$) and cross ($R_c$) validation. The reported values are averaged over five different runs.

These outliers weaken the performance measures, but they may also have a negative impact on the identification of the model. We performed thus a second round of identification, after removal from the training sets of all mutants for which $|\Delta G_{pred} - \Delta G_{exp}|$ was larger than 1.5 kcal/mol in each of the five runs performed with the same set of potentials ($G_1$, $G_2$, $G_4$, $G_5$). These consistently badly predicted mutants were however maintained in the validation sets, so as to obtain comparable results. As can be seen in the last row of Table 2, the results are even better: the correlation coefficient reaches $R_c = 0.79$ and the root mean square error drops to $\sigma_c = 0.86$.

When designing our prediction method, we excluded from the dataset the mutants with $\Delta G_0 > 5$ kcal/mol, and those involving a proline, since they are likely to induce significant structural changes (Section 2.1). Yet we estimated their $\Delta G_0$ values using the parameters that were optimized without them. The root mean square error $\sigma_i$ amounts to 1.33 kcal/mol for the 132 mutants involving a proline, which is slightly larger than the value of 1.15 kcal/mol obtained for the totality of other mutants (Table 2), but still very reasonable. The 41 mutants with $\Delta G_0 > 5$ kcal/mol are all predicted as being strongly destabilizing, but not as much as the experiments suggest: the average $\Delta G_0$ is equal to 2.3 kcal/mol, and $\sigma_i$ to 4.1 kcal/mol. Note that the values reported in Table 2 are affected by <1% when these 173 mutants are incorporated in the validation sets.

### 3.2 Biophysical significance of the parameters

The weight of the different energy terms in the folding free energy change $\Delta \Delta G_0$ [Equation (4)] are given by the $\alpha_i$ coefficients, which are functions of the solvent accessibility $A$. The shapes of most $\alpha_i(A)$ functions do not depend significantly on the training dataset, and are found to be well conserved in the different runs. This supports the idea that they have some biophysical meaning, and that they reflect the influence of the corresponding type of interactions at the protein surface and in the core. A few examples of the dependence of $\alpha_i$ on $A$ are given in Figure 2.

For the basic potentials, included in set $G_1$ (Table 1), the interpretation of the $\alpha_i$ profiles is straightforward. The potential $\Delta W_{sd}$ is computed from propensities of amino acids to be associated to torsion angle domains and thus describes interactions between amino acids close to each other in the sequence, which are at the basis of local (secondary) structure formation. As seen in Figure 2, the weight of these interactions is larger on the surface than in the core of proteins. In contrast, the weight of $\Delta W_{sd}$, which is derived from propensities of amino acids of a given type to be close to any other amino acid and is dominated by the hydrophobic effect, is larger in the core. The value of the $\alpha_i$ coefficient corresponding to $\Delta W_{sd}$, which describes specific interactions between amino acids close to each other in space, is also slightly lower on the surface. These trends are in agreement with a previous study showing that the influence of local (non-local) interactions increases (decreases) from the core to the surface (Gilis and Rooman, 1997).

The interpretation of the $\alpha_i$ functions corresponding to the coupling terms included in sets $G_2$ and $G_3$ is less straightforward since they are somehow correlated with each other and with the basic potentials from set $G_1$. Therefore, compensatory mechanisms may occur, which explain the negative values of some $\alpha_i$’s observed in Figure 2. The identification of the parameters of $G_1$ in the absence

<table>
<thead>
<tr>
<th>Potentials</th>
<th>$N$</th>
<th>$R_d$</th>
<th>$\sigma_d$/kcal/mol</th>
<th>$R_c$</th>
<th>$\sigma_c$/kcal/mol</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$, $G_2$, $G_3$, $G_4$, $G_5$</td>
<td>85/108</td>
<td>0.65/0.62</td>
<td>1.12/1.16</td>
<td>0.80/0.77</td>
<td>0.85/0.89</td>
</tr>
<tr>
<td>$G_1$, $G_2$, $G_3$, $G_4$</td>
<td>52/64</td>
<td>0.64/0.62</td>
<td>1.13/1.16</td>
<td>0.79/0.78</td>
<td>0.85/0.88</td>
</tr>
<tr>
<td>$G_1$, $G_3$, $G_5$</td>
<td>51/60</td>
<td>0.61/0.59</td>
<td>1.18/1.21</td>
<td>0.76/0.77</td>
<td>0.88/0.90</td>
</tr>
<tr>
<td>$G_1$, $G_2$, $G_4$, $G_5$</td>
<td>40/56</td>
<td>0.59/0.57</td>
<td>1.18/1.21</td>
<td>0.75/0.73</td>
<td>0.94/0.95</td>
</tr>
<tr>
<td>$G_1$, $G_2$, $G_3$, $G_4$, $G_5$</td>
<td>25/28</td>
<td>0.61/0.60</td>
<td>1.17/1.18</td>
<td>0.73/0.76</td>
<td>0.90/0.91</td>
</tr>
<tr>
<td>$G_1$, $G_2$, $G_3$, $G_4$, $G_5$</td>
<td>52/64</td>
<td>$-0.63$</td>
<td>$-1.15$</td>
<td>$-0.79$</td>
<td>$-0.86$</td>
</tr>
</tbody>
</table>
of coupling terms from $G_2$ and $G_3$ yields the parameters noted $\Delta V_{\alpha_{st}}$, $\Delta V_{\alpha_{sds}}$, and $\Delta V_{\alpha_{ind}}$, which show almost the same profile as the original parameters $\alpha_{st}$, $\alpha_{sds}$, and $\alpha_{ind}$. This attests that the overall shape of the $\alpha_i$ functions corresponding to basic potentials is modulated, but not radically modified, by the presence of coupling terms.

The $\alpha_{ind}$ parameter corresponding to the independent term is positive at small values of $A$, and tends to zero when $A$ increases. This can be explained by the fact that local side chain rearrangements that limit strong destabilization are usually possible at the surface, but are much more difficult in the core. The $\alpha_{st}$ parameter corresponding to the $\Delta V_{\alpha_{s}}$ term decreases as a function of the solvent accessibility, which correctly models the fact that the creation of a cavity is unfavourable in the core of a protein but not (as much) on the surface.

In contrast, the $\alpha_{sds}$ profile corresponding to $\Delta V_{\alpha_{sds}}$ is intriguing. It is indeed increasing as a function of increasing solvent accessibility whereas it is expected to decrease, for steric stresses resulting from the replacement of a small amino acid by a larger one should be more destabilizing in the core. This unexpected result is due to a partial compensation between the $\Delta V_{\alpha_{sds}}$ contribution to the folding free-energy change $\Delta \Delta G_F$ and that of the independent term: when the latter is not included in the model, the corresponding $\alpha_{sds}$ parameter, noted $\alpha_{sds}^\prime$, shows the expected decreasing behavior. Note that $\Delta V_{\alpha_{sds}}$ is also somewhat compensated by the independent term, as can be seen from the larger values of $\alpha_{sds}^\prime$, compared to $\alpha_{sds}$, at low $A$.

### 3.3 Comparison with other methods

Six other algorithms, previously published and available on the web, are tested against our new method for predicting stability changes upon mutation. PoPMuSiC-1.0 (Kwasigroch et al., 2002) and CUPSAT (Parthiban et al., 2006) rely on environment-dependent combinations of backbone torsion angle and distance potentials. The latter potentials are residue-based in PoPMuSiC-1.0 and atom-based in CUPSAT. Dmutant exploits atom-based distance potentials with a finite ideal gas reference state (Zhou and Zhou, 2002). Automute uses a four body, knowledge-based potential, combined with machine learning methodologies (Masso and Vaisman, 2008). I-mutant-2.0 is a support vector machine predictor, whose input vector consists of 42 elements, including the mutated and mutant amino acid, their structural environment, the temperature and pH (Capirotti et al., 2005). Eris uses a physical force-field including contributions from van der Waals forces, solvation, hydrogen bonding and backbone-dependent statistical energies (Yin et al., 2007). MuPRO (Cheng et al., 2006) was not included in our analysis, since the web server predicts only the sign of the $\Delta \Delta G$.

To assess and compare the predictive power of these methods, we built a validation set by selecting all mutants that were not included in any of the databases used to devise or test these methods. Mutations in multimeric proteins were also excluded, since it is not always clear how the different prediction servers handle such mutants, and whether the predicted stability changes correspond to the multimer or to a single monomeric unit. These two criteria resulted in a dataset of 1181 mutations. As most servers require the introduction of each mutation separately, which requires a large amount of time, a subset of 350 mutants (corresponding to 67 different proteins) was randomly selected.

For each of these 350 mutations, the predicted value of $\Delta \Delta G$ was extracted from all six web servers. In the case of our novel program PoPMuSiC-2.0, the predicted values were obtained after training of the neural network on the remaining 2298 mutants of our dataset. The correlation coefficient and root mean square error between predicted and measured $\Delta \Delta G$’s are given in Table 3, for the
complete set of 350 mutants, and after removal of 10% outliers. Note that some of these web applications failed to provide predictions for all 350 mutants. In these cases, the missing \( \Delta G_p \) values were set to 0.0 kcal/mol. To check that the overall comparison is not affected by these issues, we also computed the correlation coefficient and root mean square error on the 309 mutants for which a prediction was obtained from all servers (Table 3). A table including all \( \Delta G_p \) values set to 0.0 kcal/mol. 

In those cases, the missing \( \Delta G_p \) values were set to 0.0 kcal/mol. To check that the overall comparison is not affected by these issues, we also computed the correlation coefficient and root mean square error on the 309 mutants for which a prediction was obtained from all servers (Table 3). Indeed, \( R \) is lower than 0.5 for most methods. The highest value is 0.67, and is reached by our PoPMuSiC-2.0 program. The root mean square error is superior to 1.50 kcal/mol for most methods, and is the lowest for PoPMuSiC-2.0 (1.16 kcal/mol).

However, as discussed previously, the removal of consistently badly predicted mutants may give a more relevant view of the performances. We removed thus, for each method independently, the 10% of the mutants that lower most the correlation coefficient. Indeed, this new version of PoPMuSiC is shown to outperform other methods such as CUPSAT (Parthiban et al., 2002) and MuPRO (Cheng et al., 2006) achieved a correlation coefficient of 0.41 and a root mean square error of 1.44 kcal/mol (\( R = 0.61 \) and \( \sigma = 1.27 \) kcal/mol after removal of 10% outliers), while MuPKO (Cheng et al., 2006) achieved a correlation coefficient of 0.61 and a root mean square error of 1.27 kcal/mol (\( R = 0.61 \) and \( \sigma = 1.20 \) kcal/mol after removal of 10% outliers). Sequence-based methods appear thus as interesting but in general less performing than most structure-based methods, as expected.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of predictions(^a)</th>
<th>Complete set (350/309 mutants)(^b)</th>
<th>Exclusion of 10% outliers (315/278 mutants)(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( R ) ( \sigma ) (kcal/mol)</td>
<td>( R ) ( \sigma ) (kcal/mol)</td>
</tr>
<tr>
<td>Automate</td>
<td>315</td>
<td>0.46/0.45 1.43/1.46</td>
<td>0.670/0.68 1.15/1.20</td>
</tr>
<tr>
<td>CUPSAT</td>
<td>346</td>
<td>0.370/0.35 1.91/1.96</td>
<td>0.620/0.61 1.32/1.33</td>
</tr>
<tr>
<td>Dmutant</td>
<td>350</td>
<td>0.480/0.47 1.81/1.87</td>
<td>0.740/0.74 1.17/1.21</td>
</tr>
<tr>
<td>Iris</td>
<td>334</td>
<td>0.350/0.34 4.124/2.48</td>
<td>0.670/0.67 2.89/0.96</td>
</tr>
<tr>
<td>I-mutant-2.0</td>
<td>346</td>
<td>0.290/0.27 1.65/1.69</td>
<td>0.570/0.56 1.23/1.27</td>
</tr>
<tr>
<td>PoPMuSiC-1.0</td>
<td>350</td>
<td>0.620/0.63 1.241/2.25</td>
<td>0.770/0.77 0.98/1.00</td>
</tr>
<tr>
<td>PoPMuSiC-2.0</td>
<td>350</td>
<td>0.670/0.67 1.161/1.19</td>
<td>0.810/0.81 0.94/0.96</td>
</tr>
</tbody>
</table>

\( \sigma \) values set to 0.0 kcal/mol. The second correspond to the 309 mutants (278 after exclusion of 10% outliers) for which a \( \Delta G_p \) value is available for all predictors. Note that the outliers are selected independently for each method.

The comparative results presented in Table 3 correspond to previously published algorithms that allow the fast prediction of mutant stability changes, on the basis of the protein structure. A few other methods were developed to carry out such predictions using the sequence only. Of course, they are not expected to perform as well as structure-based predictors, but are obviously useful in a different range of applications, when no experimental or modeled structure is available. To assess the benefits of using structural information in addition to sequence, we tested two sequence-based predictors. The sequence-only version of I-mutant-2.0 (Capriotti et al., 2005) achieved a correlation coefficient of 0.30 and a root mean square error of 1.70 kcal/mol on our dataset of 350 mutants (\( R = 0.56 \) and \( \sigma = 1.27 \) kcal/mol after removal of 10% outliers), while MutK (Cheng et al., 2006) achieved a correlation coefficient of 0.41 and a root mean square error of 1.44 kcal/mol (\( R = 0.61 \) and \( \sigma = 1.27 \) kcal/mol after removal of 10% outliers). Sequence-based methods appear thus as interesting but in general less performing than most structure-based methods, as expected.

4 CONCLUSIONS

Our enhanced set of statistical potentials, along with a neural network aimed at optimizing their weights as a function of the solvent-accessibility of the mutated residues, are the two basic ingredients of our new method for predicting protein stability changes upon mutation. In contrast with the original version of PoPMuSiC (Gilis and Rooman, 2000; Kwogire et al., 2002) and other methods such as CUPSAT (Parthiban et al., 2006), our method does not rely on discrete environmental compartments in which different energy functions are used. Rather, a single energy function, including components that vary with the environment (solvent-accessibility, secondary structure), is applied to all mutations. The high predictive power of our model, assessed by a rigorous cross-validation procedure, is illustrated by a correlation coefficient of 0.79 between the predicted and measured \( \Delta G \) values, and a root mean square error of 0.86 kcal/mol, on 90% of the data. In addition, this new version of PoPMuSiC is shown to outperform...
all (to our knowledge) previously published and freely available structure-based predictors, on an independent set of more than 300 mutants.

Despite the use of a neural network to optimize the weighting functions of the different potentials, our predictive model is very different from a pure black box model. Indeed, the parameters identified by the neural network define sigmoid functions describing the transition between two different environments, in which different types of interactions dominate. The overall shapes of these sigmoid functions were found to be very stable with respect to the composition of the training set, and may thus be considered as having a genuine biophysical significance. In particular, our results confirm previous observations that local interactions are more influential on the surface of the protein and tertiary interactions dominate in the core (Gilis and Rooman, 1997). They also indicate that changing the amino acid size upon mutation tends to be destabilizing, especially in the core, as this creates either stress or cavities, and that all mutations are on the average more destabilizing in the core than at the surface.

Packaging defects caused by mutations in the core cannot be correctly taken into account by potentials derived from a database of wild-type proteins, since these proteins are usually very well packed. The volume terms included in our energy function, along with the independent term, provide a very coarse description of these defects but still enhance substantially the predictive power of our method. Among the further improvements that should be introduced in PoPMuSiC, one of the priorities is certainly a better description of the impact of packing defects. The modeling of possible local structural rearrangements of the side-chains or backbone upon mutation should improve the performances of our method, and allow us to consider multiple mutations in a systematic way. However, taking into account structural flexibility is usually very hungry for computing power, while the ability of our current algorithm to predict within a minute (on a contemporary desktop computer) the stability changes resulting from all possible mutations in an average-sized protein is definitely one of its advantages.

In summary, our results suggest that PoPMuSiC-2.0 is—at present—the most reliable method for predicting stability changes of mutant proteins, as compared to the other available structure-based predictors that are able to test rapidly all possible mutations in a given protein. We believe that every rational design of modified proteins would benefit from the use of PoPMuSiC-2.0 in a first stage, so as to identify a small set of mutations likely to present the desired (de)stabilization properties. These mutations are then to be studied further, either by more time-consuming computational methods, or directly by experimental means (see for example Cahiria et al., 2007; Gilis et al., 2005).

ACKNOWLEDGEMENTS

The authors thank Jean Marc Kwasirogro for his contribution to the web server.

Funding: Belgian State Science Policy Office through an Interuniversity Attraction Poles Programme (DYSCO); the Belgian Fund for Scientific Research (FRS) (FREc project, FRIA grants to A.G and B.F.); the Brussels Region (TheraVip project); the Walloon Region and BioXpr bioinformatics company (First-Postdoc grant to Y.D.). M.R. is Research Director at the FRS.

Conflict of Interest: none declared.

REFERENCES


