



A neural-network-based method for predicting protein stability changes upon single point mutations

Emidio Capriotti, Piero Fariselli* and Rita Casadio

Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Imerio 42, 40126 Bologna, Italy

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: One important requirement for protein design is to be able to predict changes of protein stability upon mutation. Different methods addressing this task have been described and their performance tested considering global linear correlation between predicted and experimental data. Neither is direct statistical evaluation of their prediction performance available, nor is a direct comparison among different approaches possible. Recently, a significant database of thermodynamic data on protein stability changes upon single point mutation has been generated (ProTherm). This allows the application of machine learning techniques to predicting free energy stability changes upon mutation starting from the protein sequence.

Results: In this paper, we present a neural-network-based method to predict if a given mutation increases or decreases the protein thermodynamic stability with respect to the native structure. Using a dataset consisting of 1615 mutations, our predictor correctly classifies >80% of the mutations in the database. On the same task and using the same data, our predictor performs better than other methods available on the Web. Moreover, when our system is coupled with energy-based methods, the joint prediction accuracy increases up to 90%, suggesting that it can be used to increase also the performance of pre-existing methods, and generally to improve protein design strategies.

Availability: The server is under construction and will be available at <http://www.biocomp.unibo.it>

Contact: piero.fariselli@unibo.it

1 INTRODUCTION

The comprehension of the rules that govern protein stability is one of the long-term goals of protein structure analysis (Daggett and Fersht, 2003). This would also help in protein designing. For these reasons, so far different methods have been described to predict stability changes observed upon residue substitution in the original protein sequence. They are mainly based on the development of different energy

functions, suited to compute the stability free energy changes due to substitution of one residue at a time in the protein sequence (mutation).

Force fields previously described can be grouped into three major categories depending on the energy functions (Guerois *et al.*, 2002): (1) physically effective energy functions (Prevost *et al.*, 1991; Pitera and Kollman, 2000); (2) statistical potential energy functions (Gilis and Rooman, 1997; Kwasigroch *et al.*, 2002); and (3) empirically defined energy functions (Funahashi *et al.*, 2001; Guerois *et al.*, 2002; Zhou and Zhou, 2002). The major drawback of methods based on physically effective energy functions (Prevost *et al.*, 1991; Pitera and Kollman, 2000) is that they are computationally intensive so that their usage is nearly prohibitive for applications and/or screening on a large scale (Guerois *et al.*, 2002). For this reason, only the last two types of energy functions can practically be applied for evaluating protein stability on a large dataset (Gilis and Rooman, 1997; Funahashi *et al.*, 2001; Guerois *et al.*, 2002; Zhou and Zhou, 2002).

Energy-based methods routinely compute the relative stability changes upon protein mutation ($\Delta\Delta G$). Their accuracy towards the experimental database is then evaluated by considering the correlation between the predicted data against the experimental ones. Global correlation can be as high as 95% depending on the selected mutation database (Guerois *et al.*, 2002). In many instances, for modelling protein stability, it would be convenient to know the reliability of a method and the statistical scores associated with any predicted values of free energy stability change upon single point mutation (single residue substitution) in a given protein chain. Indeed any predicted $\Delta\Delta G$ consists of a numerical absolute value (the amount of stability change) and of a sign, namely, the prediction of the direction in which the mutation will shift the stability of the protein (either towards a positive or negative $\Delta\Delta G$ value, corresponding to an increase or decrease in stability, respectively). The correct prediction of the direction of the stability change is therefore more relevant than its absolute value for the problem at hand. A linear correlation between predicted and experimental data, although exceedingly good,

*To whom correspondence should be addressed.

does not provide direct indication of the correctness of the $\Delta\Delta G$ sign.

Other relevant thermodynamic parameters in mutagenesis are experimental conditions, such as pH and temperature (Gromiha *et al.*, 2000). In this respect, energy-based methods need to fit these parameters assuming that the mutations are carried out at physiological conditions. This limitation can be also overpassed using machine learning approaches.

Thanks to the availability of a large database of thermodynamic data of mutated proteins (Gromiha *et al.*, 2000), it is now possible to apply machine learning techniques to the relevant problem of predicting protein stability changes upon single point mutations. We reported previously how protein stability from sequence alone (Casadio *et al.*, 1995) should be predicted. Here, we develop a neural-network-based method to predict if a given mutation increases or decreases the native protein stability, taking also into account the three-dimensional structure of the protein and focusing specifically on the prediction of the sign of the $\Delta\Delta G$ value. We show that it reaches accuracy as high as 81% and in the specific task of predicting the $\Delta\Delta G$ sign, it performs better than the available energy-based methods. Moreover, we show that when it is coupled with energy-based predictors the accuracy of the joint method significantly increases with respect to that of the original ones.

2 SYSTEMS AND METHODS

2.1 The protein database

Our dataset is derived from the current release (July 2003) of the Thermodynamic Database for Proteins and Mutants [ProTherm by Gromiha *et al.* (2000)]. We considered two datasets: the first for training/testing our neural network system (S1615), and the second (a subset of the first), to be used in a testing phase with cross-validation procedure for comparison with other available predictors, considering mutations only at physiological conditions (S388).

The dataset of proteins was extracted from ProTherm, with the following constraints:

- (1) the $\Delta\Delta G$ value was experimentally detected and reported in the database;
- (2) the protein structure is known with atomic resolution [and deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2000)];
- (3) the data are relative to single mutations (no multiple mutations have been taken into account).

After this filtering procedure, we end up with a dataset consisting of 1615 (S1615) different single mutations obtained from 42 different protein sequences, known with atomic resolution. The second set, containing only experiments performed at physiological conditions, namely temperatures in the range of 20–40°C, and pH values in the range of 6–8, is a subset of

S1615 and consists of 388 mutations (S388) from 17 different protein sequences. This was generated in order to compare our method with those that consider only the effects of mutations under physiological conditions. The final S388 and S1615 sets are available at <http://www.biocomp.unibo.it/piero/ddgp/>

2.2 The neural-network-based predictor

Our task is to predict whether a given mutation increases or decreases the protein stability, without predicting the exact $\Delta\Delta G$ value. In this respect, the task can be cast as a classification problem for the protein upon mutation. To address this issue we use standard feed-forward neural networks, with the back-propagation algorithm as a learning procedure. The network architecture consists of a one-layer perceptron with two hidden nodes and one output node, that codifies for the increased protein stability ($\Delta\Delta G > 0$, desired output set to 1) or for the destabilising mutation ($\Delta\Delta G < 0$, desired output set to 0). The decision threshold is set equal to 0.5.

In order to highlight the major features responsible for protein stability changes upon mutation, we introduced different input encodings, which are listed below (in order of increasing complexity).

N1 consists of 22 input neurons, 2 of which account respectively for the temperature and the pH at which the stability of the mutated protein was measured, while the other 20 nodes explicitly define the mutation. These 20 elements code for the corresponding 20 amino acid residues, so that given a residue mutation, we set to -1 the element corresponding to the deleted residue and to 1 the new introduced residue (all the remaining elements are kept equal to 0).

N2 adds to the N1 input one more neuron that accounts for the relative accessibility of the mutated residue computed with the DSSP program (Kabsch and Sander, 1983). This information can be very profitable, since it has been shown that residues at the surface of a protein routinely contribute less to the protein stability than those in the core (Guerois *et al.*, 2002).

Finally, N3 uses 20 more input neurons (43 in total) encoding the three-dimensional residue environment (again the 20 neurons represent the 20 residue types). Each of this input neuron is provided with the number of the encoded residue type, found inside a sphere of radius T , which is centred in the mutated residue and is taken from the corresponding protein structure. More formally, given a residue in the sequence position i of coordinate $r(i)$, the element a of the input vector V (of 20 components) is computed as

$$V(a) = \sum_j \delta[\text{type}(j), \text{type}(a)] \rho[r(i), r(j), R] \quad (1)$$

where j spans the protein length; $\delta[\text{type}(j), \text{type}(a)]$ is set equal to 1 only when the residue in position j is equal to type a ; $\rho[r(i), r(j), R]$ is also set to 1 only if the Euclidean distance between $r(i)$ and $r(j)$ is lower than the threshold R (the sphere radius). To compute Equation (1), we use a full-atom representation of the protein (without hydrogen atoms)

(Fariselli and Casadio, 1999) and different values of the sphere radius (R), to include the effect of different contact residues (Pollastri *et al.*, 2002).

2.3 Scoring the performance

All the results obtained with our neural network systems are evaluated using a cross-validation procedure on the S1615 dataset.

Several measures of accuracy are routinely used. For sake of completeness here we review the ones adopted in this paper. The efficiency of the predictor is scored using the statistical indexes defined in the following. The overall accuracy is

$$Q2 = p/N, \quad (2)$$

where p is the total number of correctly predicted residues and N is the total number of residues.

The correlation coefficient C is defined as

$$C(s) = \frac{[p(s)n(s) - u(s)o(s)]}{D}, \quad (3)$$

where D is the normalization factor

$$D = \{[p(s)+u(s)][p(s)+o(s)][n(s)+u(s)][n(s)+o(s)]\}^{1/2} \quad (4)$$

for each class s (+ and -, for positive and negative $\Delta\Delta G$ values, respectively); $p(s)$ and $n(s)$ are the total number of correct predictions and correctly rejected assignments, respectively, and $u(s)$ and $o(s)$ are the numbers of under and over predictions.

The coverage for each discriminated structure s is evaluated as

$$Q(s) = p(s) / [p(s) + u(s)], \quad (5)$$

where $p(s)$ and $u(s)$ are the same as in Equation (3). The probability of correct predictions $P(s)$ (or accuracy for s) is computed as

$$P(s) = \frac{p(s)}{[p(s) + o(s)]}, \quad (6)$$

where $p(s)$ and $o(s)$ are the same as in Equation (3) (ranging from 1 to 0).

Finally, it is very important to assign a reliability score to each network prediction. With one output neural network this is obtained by computing

$$\text{Rel}(i) = 20 * \text{abs}[O(i) - 0.5], \quad (7)$$

where $O(i)$ is the network output.

3 RESULTS AND DISCUSSION

3.1 The effect of the local spatial environment

In the database, mutations per protein are related to the protein stability energy change ($\Delta\Delta G$). Therefore, a first generalization that the network can perform depending on the pH and temperature is how relevant to the protein stability the mutation type is, independent of the protein context.

Table 1. Cross-validation performance of the different neural networks on S1615

Method	Q2	P(+)	Q(+)	P(-)	Q(-)	C
N1	0.74	0.59	0.23	0.76	0.94	0.24
N2	0.75	0.57	0.45	0.80	0.87	0.34
N3	0.79	0.63	0.55	0.83	0.88	0.45

+ and -: the index is evaluated for positive and negative signs of protein energy stability change, respectively; for the definition of the different indexes see the System and Methods section.

Network N1 indeed generalizes over this information, and its accuracy depends only on the mutation itself. From Table 1, we can note that N1 can correctly predict 74% of the dataset, indicating that the pairwise mutation that is given as input to the network carries already information relevant to correctly predict the $\Delta\Delta G$ sign. When the network input includes the information relative to the solvent accessibility value of the mutated residue (N2), the prediction slightly improves, as indicated by the increase of the correlation coefficient value listed in Table 1 (from 0.24 to 0.34).

This finding is in agreement with the common view that protein stability would be more dependent on buried residues than on solvent-exposed ones (Guerois *et al.*, 2002; Zhou and Zhou, 2004).

With N3 the three-dimensional environment of the mutated residue is taken into account. In this case (Fig. 1), the local environment of the residue undergoing mutation is evaluated by considering all the contact residues within different contact distances. We adopt an all-atom representation of the protein and we started with a local three-dimensional environment computed by setting the sphere radius equal to 4.5 Å. This threshold identifies the largest shell that does not allow a water molecule to be inserted between two residues (Fariselli and Casadio, 1999). It is evident that this information promotes an increase of the accuracy up to 79% and of the correlation coefficient up to 0.45, 20 percentage points higher than that obtained using the level of knowledge encoded in N1 (results obtained with N3 in Table 1).

In Table 2, we investigate the effect of enlarging the three-dimensional environment by increasing the sphere radius from 4.5 to 12.0 Å [R in Equation (1)]. It is worth noticing that there is an optimal radius value corresponding to 9.0 Å. In this case, N3 can correctly predict >80% of the dataset with a correlation coefficient of 0.49.

As in the case of other neural-network-based predictors, the overall Q2 accuracy can be computed as function of the reliability index (Rel) (Fig. 2). This identifies a relationship between the reliability value and the predictor accuracy, that may help in selecting which mutations are more suited to increase or decrease protein stability in a rational computer-aided protein design.

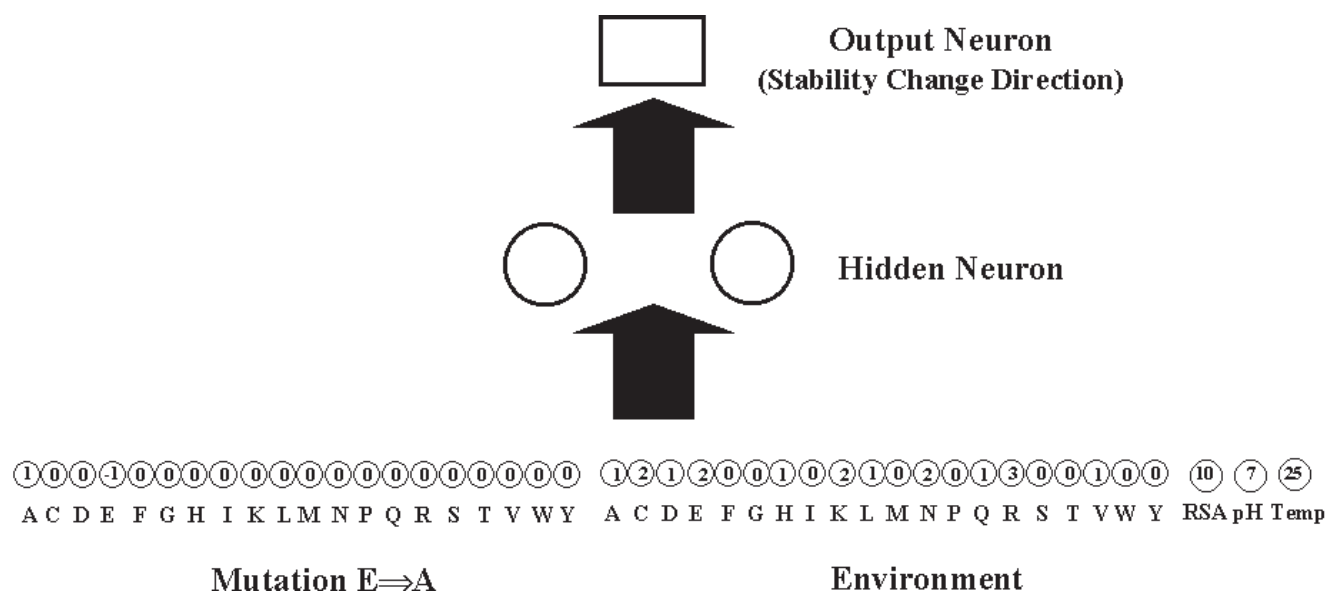


Fig. 1. N3 input encoding includes the mutation type (20 element vector); the local spatial environment (20 element vector) as defined in Equation (1); the relative solvent accessibility value (RSA); and the experimental conditions at which the mutation was carried out [pH and the temperature (Temp)].

Table 2. Cross-validation performance of N3 as a function of different protein environments centred on the mutated residue

Method	$R(\text{Å})$	Q2	P(+)	Q(+)	P(-)	Q(-)	C
N3-4.5	4.5	0.79	0.63	0.55	0.83	0.88	0.45
N3-6.0	6.0	0.79	0.63	0.57	0.84	0.87	0.46
N3-9.0	9.0	0.81	0.71	0.52	0.83	0.91	0.49
N3-12.0	12.0	0.79	0.63	0.59	0.84	0.87	0.47

R = radius [Equation (1)]. For notation see Table 1.

Our approach presently does not take into account the change in environment (local structure) that can occur upon protein mutation, since the environment description is essentially derived from the native protein, before mutation. Nevertheless this approximation is sufficient to increase the predictive capability of the system, indicating that the network is able to satisfactorily relate the mutation type, the structural environment in the native protein and the correspondent $\Delta\Delta G$ sign upon mutation.

3.2 Comparison with other methods

Differently from our approach, all the methods available use the molecular protein structure at different level of representation for predicting the $\Delta\Delta G$ values after the protein mutation (Gilis and Rooman, 1997; Funahashi *et al.*, 2001; Guerois *et al.*, 2002; Kwasigroch *et al.*, 2002; Zhou and Zhou, 2002). In this respect a direct comparison between other methods and ours is not feasible.

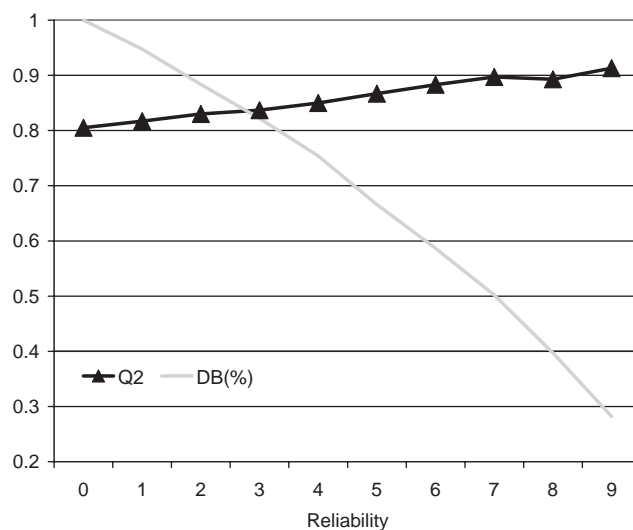


Fig. 2. Q2 accuracy of neural network (N3-9.0) as a function of the reliability index (Rel) of the prediction [Equation(7)]. DB% is the percentage of the dataset with Rel values higher or equal to a given threshold.

However, since the feature that our predictor highlights is the direction (sign) of the stability change, we compare the N3-9.0 predictions with those computed using the three presently available web servers, implementing some of the energy-functions-based methods: FOLDX (Guerois *et al.*, 2002); DFIRE (Zhou and Zhou, 2002) and PoPMuSiC (Gilis and Rooman, 1997; Kwasigroch *et al.*, 2002). All the servers are compared in the task of predicting

Table 3. Comparison of neural network with other methods on S388

Method	Q2	P(+)	Q(+)	P(-)	Q(-)	C
FOLDX ^a	0.75	0.26	0.56	0.93	0.78	0.25
DFIRE ^b	0.68	0.18	0.44	0.90	0.71	0.11
PoPMuSiC ^c	0.85	0.33	0.25	0.90	0.93	0.20
N3-9.0	0.87	0.44	0.21	0.90	0.96	0.25

For notation see Table 1.

^a<http://fold-x.embl-heidelberg.de>

^b<http://phyzz4.med.buffalo.edu/hzhou/dmutation.html>

^c<http://babylone.ulb.ac.be/popmusic/>.

if a given mutation increases or decreases the protein stability (using only the sign of the prediction).

In Table 3, we show the results obtained with N3-9.0 (testing on Subset S388, that contains only mutations at physiological conditions and in order to compare with the web predictors that do not take into account the pH and temperature dependence). All the methods, with exception of our predictor and PoPMuSiC (that it is not trained on $\Delta\Delta G$ values) were used without considering if a given mutation had been included or not in the training set. The results reported in Table 3 show that N3-9.0, for this specific task, performs similarly or better than the other methods. This indicates that the N3 neural network is capable of extracting features from the data that are qualitatively comparable to those explicitly introduced using energy functions.

3.3 Analysis of the prediction

Focusing on N3-9.0 performance, we analyse the quality and the major characteristics of the N3 predictions. In Figure 3, we report the Q2 accuracy values as a function of the experimental $\Delta\Delta G$ absolute values derived from the database (the larger the value the greater the stability change). It is worth noticing that even though no information about the experimental $\Delta\Delta G$ values was provided to the network, the accuracy is lower for the smaller changes and higher for the larger ones. This finding suggests that the approximation of our approach (including the relevance of the protein environment) holds particularly for large changes of protein stability, that are also the most relevant to be predicted for the *in silico* molecular design.

In Figure 4, N3-9.0 accuracy (Q2) is plotted as a function of the solvent accessibility value of the mutated residue. Apparently, the sign of protein stability change upon exposed residue mutation (with a solvent accessibility value > 60%) is on average more difficult to be predicted than that of less-exposed residues. Our finding is in agreement with previous observations (Guerois *et al.*, 2002; Zhou and Zhou, 2004) indicating that mutated residues at the surface behave differently from those buried and their effect on protein stability is more difficult to predict. At variance, PoPMuSiC obtains better results for mutation of highly exposed residues than for those

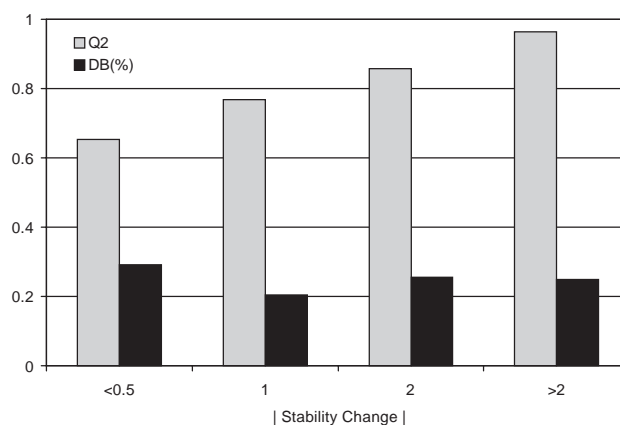


Fig. 3. Q2 accuracy of neural network (N3-9.0) as a function of the absolute value of protein stability changes upon mutation ($|\text{Stability Change}|$ = is the absolute value of the stability change upon mutation ($\Delta\Delta G$). DB% is the percentage of the dataset that falls in a specific range of $|\text{Stability Change}|$.

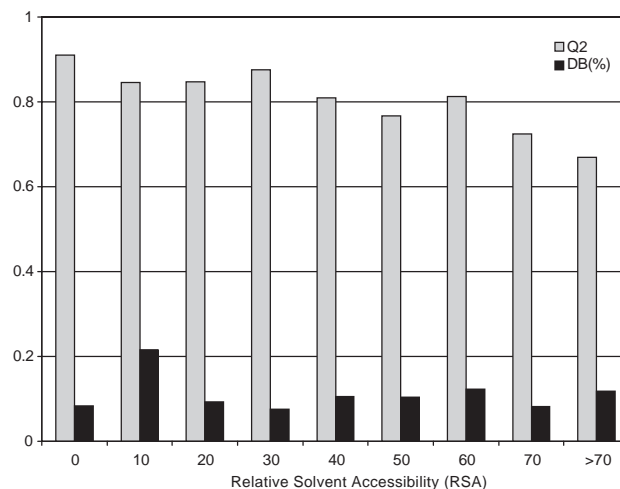


Fig. 4. Q2 accuracy of the neural network (N3-9.0) as a function of the relative accessibility value of the mutated residue. DB% is the percentage of the data with a given relative solvent accessibility value.

of buried residues (Gilis and Rooman, 1997; Guerois *et al.*, 2002).

In Table 4, the accuracy on N3-9.0 is also shown as a function of the chemical-physical type of residue mutation (from charged, polar and apolar residues to charged, polar and apolar, respectively).

Apparently it is more difficult to predict the protein stability change in the case of the charged/charged or polar/charged mutation than in other cases (as indicated by lower accuracy values). Since the frequency of occurrence of polar and charged residues is on average higher on the protein solvent accessibility surface, the results shown in Table 4 are again consistent with the observation (above) that our predictor is

Table 4. Q2 accuracy as a function of the residue mutation type

Native\new	Charged	Polar	Apolar
Charged	0.62 (4%)	0.77 (8%)	0.72 (9%)
Polar	0.69 (6%)	0.82 (10%)	0.77 (17%)
Apolar	0.75 (3%)	0.92 (12%)	0.87 (31%)

Each cell represents a particular type of mutation classified according to chemico-physical properties. Rows account for the wild-type residue, while the column positions define the new residues in the mutant proteins (new). In parentheses, the relative frequency in the S1615 dataset of a given residue type is shown.

Table 5. Accuracy of joint-methods on subsets of S388

Method	Agree	Q2	P(+)	Q(+)	P(-)	Q(-)	C
N3-9.0+FOLDX ^a	72%	0.93	0.88	0.28	0.93	0.99	0.47
N3-9.0+DFIRE ^b	69%	0.90	0.36	0.16	0.92	0.97	0.19
N3-9.0+PoPMSiC ^c	86%	0.91	0.67	0.07	0.92	0.99	0.19

Agree = percentage of cases on which both methods agree on the predicted sign of protein stability change. The statistical indices are evaluated on the agreement subset. For notation see Table 3.

more accurate in predicting residues characterized by low solvent accessibility values.

3.4 Combining the neural network with the energy-based predictors

N3-9.0 is then used to select predictions of the existing energy-based methods with the criterion that only $\Delta\Delta G$ values with the same sign of that predicted by the network are accepted and included in a subset of S388 named Agreement in Table 5. This procedure is applied by combining the network acting as a filter to the methods scored in Table 3, and the predictions are then re-scored (Table 5). Apparently the joint predictions improve the overall performance of the energy-based methods; for example a Q2 of 93% is obtained in the case of FOLDX+N3-9.0, with a probability value of 88% [P(+)] for correctly assigning a predicted mutation that increases protein stability.

Our results, altogether, suggest that a neural-network-based system, also combined with energy-based methods, can be used to properly address the problem of detecting protein stability change upon mutation and to make a step forward in the prediction task.

ACKNOWLEDGEMENTS

This work was supported by the following grants delivered to R.C.: 'Hydrolases from Thermophiles: Structure, Function and Homologous and Heterologous Expression' and the project 'Development and implementation of algorithms for predicting protein structure' of the Ministero della Università

e della Ricerca Scientifica e Tecnologica (MURST), Molecular Genetics and Functional Genomics, both of the Italian Centro Nazionale delle Ricerche (CNR), and a PNR 2001–2003 (FIRB art.8) project on Postgenomics. E.C. is supported by a grant delivered by a FISIR2002 project to R.C.

REFERENCES

- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Casadio, R., Compiani, M., Fariselli, P. and Vivarelli, F. (1995) Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 81–88.
- Daggett, V. and Fersht, A.R. (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.*, **28**, 18–25.
- Fariselli, P. and Casadio, R. (1999) Neural network based predictor of residue contacts in proteins. *Protein Eng.*, **12**, 15–21.
- Funahashi, J., Takano, K. and Yutani, K. (2001) Are the parameters of various stabilization factors estimated from mutant human lysozymes compatible with other proteins? *Protein Eng.*, **14**, 127–134.
- Gilis, D. and Rومان, M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **272**, 276–290.
- Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., Prabakaran, P. and Sarai, A. (2000) ProTherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **28**, 283–285.
- Guerois, R., Nielsen, J.E. and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern of hydrogen-bonded and geometrical features. *Bio polymers*, **22**, 2577–2637.
- Kwasigroch, J.M., Gilis, D., Dehouck, Y. and Rومان, M. (2002) PoP-MuSiC, rationally designing point mutations in protein structures. *Bioinformatics*, **18**, 1701–1702.
- Pitera, J.W. and Kollman, P.A. (2000) Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. *Proteins*, **41**, 385–397.
- Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.
- Prevost, M., Wodak, S.J., Tidor, B. and Karplus, M. (1991) Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96-Ala mutation in barnase. *Proc. Natl Acad. Sci. USA*, **88**, 10880–10884.
- Zhou, H. and 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.
- Zhou, H. and Zhou, Y. (2004) Quantifying the effect of burial of amino acid residues on protein stability. *Proteins*, **54**, 315–322.