Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Neural Networks to predict protein stability changes upon mutation

A. Grosfils*, Y. Dehouck, D. Gilis**, M. Rooman**, Ph. Bogaerts***

** Groupe de Biomodélisation et Bioprocédés, Université Libre de Bruxelles, 1050 Brussels, Belgium (e-mail: agrosfils/Philippe.Bogaerts@ulb.ac.be)*
*** Unité de Bioinformatique génomique et structurale, Université Libre de Bruxelles, 1050 Brussels, Belgium (e-mail: ydehouck/dgilis/mrooman@ulb.ac.be).*

Abstract: Black box modelling is used here to improve the performances of the PoPMuSiC program that predicts protein stability changes caused by single-site mutations. For that purpose previously developed statistical energy functions are exploited, which are based on a formalism that highlights the coupling between 4 different protein descriptors (sequence, distance, torsion angles and solvent-accessibility), as well as the volume variation of the mutated amino acid. As the importance of the different types of interactions may depend on the protein region, the stability change is expressed as a linear combination of these energetic functions, whose proportionality coefficients vary with the solvent-accessibility of the mutated residue. Two alternative structures are considered for these coefficients: a Radial Basis Function network, and a MultiLayer Perceptron with sigmoid nodes. These two structures are identified, leading to an improvement of the predictive capabilities of PoPMuSiC, and are discussed in terms of their biophysical interpretation.

1. INTRODUCTION

Proteins are the most abundant biological macromolecules. They are essential parts of all living organisms and participate in every process within cells (Creighton, 1993, Lehninger *et al*., 1993); they provide structure, catalyze cellular reactions, protect organisms against injury, transport specific molecules like oxygen and carry out a multitude of other tasks. Proteins are linear polymers composed of 20 amino acids, which differ from each other by the succession (sequence) of amino acids. This sequence determines the (generally) unique three-dimensional structure which allows the protein to carry out its biological function. The stability of this structure is thermodynamically characterized by its folding free energy, that is, the difference in free energy between folded and denatured states.

Proteins are largely used in the industrial world where their properties are exploited as well for the design of vaccines as in the agro-alimentary field. However, it can be interesting to tune certain physicochemical or biological properties through the substitution of amino acids by others. Such mutations may, for instance, increase the protein's solubility, or maintain its activity under unusual pH or temperature conditions. Whatever the modified property, one has to check if the considered mutations do not alter the protein structure and stability too much. Indeed, structure and stability deteriorations can lead to the loss of the main protein function. Of course, the experimental determination of the change in folding free energy upon mutation leads to the most reliable information. However, it is time consuming and thus cannot be used to test all possible mutations in a protein. This is why predictive methods are developed.

1.1 Stability prediction methods

Only a few theoretical methods have been developed to estimate stability changes caused by mutations. The earliest ones are based on detailed atomic models combined with semi-empirical potentials (Basch et al, 1987; Tidor and Karplus, 1991). However, such methods are computer time-consuming and cannot be used to test a large set of mutations. To avoid this problem, faster methods have been developed. They rely on rougher descriptions of the protein structure and approximate energy functions (Muñoz and Serrano, 1994, Miyazawa and Jernigan, 1994, Sippl, 1995), or bind stability changes to shape, flexibility and volume of the substituted and neighbouring amino acids (van Gunsteren and Mark, 1992, Shortle et al., 1990, Eriksson et al., 1992). The performances of these methods are reasonably good. However, the tests were restricted to a small number of mutations in a single protein, usually even at a single site, and it became manifest that these methods are not general enough to predict stability changes caused by mutations on any point in any protein.

To propose a more general prediction model, some of us previously developed the PoPMuSiC (Prediction Of Protein Mutation Stability Changes – http://babylone.ulb.ac.be) program (Gilis and Rooman, 2000; Kwasigroch et al., 2002). The energetic functions used by PoPMuSiC are statistical potentials derived from a database of known protein structures. More precisely, two types of potentials are considered: distance potentials, describing interactions between amino acids spatially close to each other, and torsion potentials, reflecting the preferences for specific local arrangements of the protein chain. The major novelty of this method is that the specific environment of the mutated amino

acid is taken into account, in order to acknowledge the fact that different types of interactions may dominate in the core and on the surface of proteins. The change in stability is thus estimated by three linear combinations of these two potentials, corresponding to three ranges of solvent accessibility of the mutated residue. PoPMuSiC has been tested on experimentally studied mutations, introduced in various environments of seven different proteins and a synthetic peptide. The correlation coefficient between predicted and measured stability changes is quite good: between 0.8 and 0.87 (according to the range of solvent accessibility) on 279 out of 296 mutants.

Since then, other predictive methods have been developed. Some are based on an approach similar to that of PoPMuSiC (Parthiban *et al*., 2006), others on neural networks (Capriotti *et al*., 2005), on decision trees (Huang *et al*., 2007), or on empirical potentials that describe the physical interactions contributing to protein stability (Guerois *et al*., 2002). PoPMuSiC still remains competitive, but it suffers from some limitations. In particular, no linear combination of the considered potentials allows to evaluate the mutations of amino acids with a solvent accessibility between 40% and 50%, and the number of experimentally characterized mutants used to validate PoPMuSiC is relatively small with respect to the currently available databases. The aim of our work is to overcome these limitations and to improve the predictive capabilities of PoPMuSiC.

2. ENERGETIC FUNCTIONS

The PoPMuSiC program relies on a set of energetic functions that describe the different interactions contributing to protein stability. These functions are statistical potentials, extracted from a database of known protein structures. Such potentials are widely used in theoretical protein studies, as they present the advantage of being easily adapted to simplified protein representations. We propose to exploit the new formalism of derivation of statistical potentials, recently developed by some of us (Dehouck *et al*., 2006), and to use these new energetic functions to improve the performances of PoPMuSiC. We first define the sequence and structure features considered in the representation of the proteins.

2.1 Simplified protein representation

The sequence of a protein is described by the *nature of the amino acid* at each position i, s_i. Its three-dimensional structure is represented by several structural descriptors. Firstly, each amino acid has several degrees of freedom, corresponding to rotations around the chemical bonds of the protein backbone. The conformation of an amino acid at position i is thus described by the *domain of torsion*, t_i, defined by its backbone torsion angles. Seven discrete values of t are considered, corresponding to specific local organisations of the protein chain (Rooman *et al*., 1991). Secondly, the *spatial distance* that separates two amino acids, at positions i and j, in the folded structure of the protein, is referred to as d_{ij}. Note that this distance is computed between the geometric centres of the amino acids' side chains, and distributed in bins of 0.2Å width. Finally, the solvent-accessibility of the amino acid at position i, a_i, 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 a are considered (0-5%, 5-15%, 15-30%, 30-50%, and 50-100%).

2.2 Statistical potentials

A form commonly used for statistical potentials derived from a set of protein structures is:

$$\Delta W(c_1,c_2) = -kT \log \frac{P(c_1,c_2)}{P(c_1)P(c_2)} \quad , \tag{1}$$

where c_1 and c_2 are sequence or structure descriptors (i.e. s_i, t_i, a_i, or d_{ij}) of the same amino acid, or of two neighboring ones, and P are their relative frequencies of occurrence in a large dataset of protein structures. Summing $\Delta W(c_1,c_2)$ over all (c_1,c_2) couples in a given protein yields an estimation of this protein's folding free energy.

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):

$$\Delta W(c_1,c_2,...,c_n) = -kT \log \prod_{k=n,n-2,n-4,...} \frac{\prod_{\substack{i_1,...,i_k=1 \\ i_1<...<i_k}}^{n} P(c_{i_1},c_{i_2},...,c_{i_k})}{\prod_{\substack{i_1,...,i_{k-1}=1 \\ i_1<...<i_{k-1}}}^{n} P(c_{i_1},c_{i_2},...,c_{i_{k-1}})} \quad , \tag{2}$$

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. Note that two major classes of potentials are considered. Local potentials describe the correlations between descriptors attached to amino acids close to each other along the sequence; ΔW_{st}, for instance, reflects the influence of the nature of an amino acid (s_i) on the conformation (t_j) of a neighbouring amino acid. Distance potentials describe the correlations between descriptors attached to amino acids close to each other in space; ΔW_{std}, for instance, reflects the propensity of an amino acid of nature s_i, in a conformation t_i, to be separated from another amino acid (whatever its nature and conformation) by a spatial distance d_{ij}.

2.3 Volume variations

Besides these energetic functions, another parameter influencing the mutant stability is envisaged: the volume difference ΔV between the wild-type and the mutant amino acids. If the mutant amino acid is smaller, a cavity is created, which usually destabilizes the protein (Eriksson et al., 1992). On the other hand, if the mutant is larger, it induces a stress in the structure, which may also have a destabilizing effect. As the amplitude of these effects are not necessarily similar, we consider them separately, by introducing ΔV_1 and ΔV_2 :

$$\Delta V_1 = \begin{cases} 0 & if\ \Delta V < 0 \\ \Delta V & if\ \Delta V \geq 0 \end{cases}, \quad \Delta V_2 = \begin{cases} \Delta V & if\ \Delta V < 0 \\ 0 & if\ \Delta V \geq 0 \end{cases} . \quad (3)$$

Table 1. Selection of statistical potentials

Local Potentials	ΔW_{st}, ΔW_{stt}, ΔW_{sst}, ΔW_{sttt}, ΔW_{sstt}, ΔW_{ssst}, ΔW_{as}, ΔW_{aas}, ΔW_{ass}, ΔW_{aaas}, ΔW_{aass}, ΔW_{asss}, ΔW_{ast}, ΔW_{aast}, ΔW_{asst}, ΔW_{astt}
Distance Potentials	ΔW_{sd}, ΔW_{asd}, ΔW_{std}, ΔW_{astd}, ΔW_{sds}, ΔW_{asdas}, ΔW_{stdst}, $\Delta W_{astdast}$

3. MODELLING PROTEIN STABILITY CHANGES WITH ARTIFICIAL NEURAL NETWORKS

PoPMuSiC models free energy changes with three linear combinations of potentials corresponding to three solvent-accessibility ranges, $A \leq 20\%$, $20 < A \leq 40\%$, and $A \geq 50\%$, where A is the solvent accessibility of the mutated amino acid. These relationships highlight the dominating influence of different types of interactions at the protein surface and in the protein core. However, no model exists for the accessibility range $40 \leq A < 50\%$. Besides the use of the new statistical potentials described above, a possible improvement is to replace the three linear combinations by only one, valid for all values of A. Its expression would be

$$\Delta\Delta G_P = \sum_{i=1}^{24} \alpha_i(A)\Delta\Delta W_i + \alpha_{25}(A)\Delta V_1 + \alpha_{26}(A)\Delta V_2 + \alpha_{27}(A), \quad (4)$$

where $\Delta\Delta G_P$ is the predicted change in folding free energy upon mutation, $\Delta\Delta W_i$ represent the 24 potentials listed in Table 1, and $\alpha_i(A)$ are the proportionality coefficients which vary with the solvent accessibility of the mutated residue. To identify the $\alpha_i(A)$ functions, we consider two different neural network structures: a radial basis function network and a multilayer perceptron.

3.1 Radial Basis Function network (RBF)

It consists in a neural network with one input node, the solvent accessibility of the mutated residue (A), three hidden neurons whose Gaussians cover the whole range of solvent accessibility and a linear output layer which provides the 27 $\alpha_i(A)$ (Fig. 1). Larger amounts of hidden neurons have been tested but they lead to redundant information. The mathematical expression of this network is

$$\alpha_i(A) = \sum_{j=1}^{3} w_{ji}\, e^{-\frac{(A-C_j)^2}{r_j^2}} + b_i, \quad (5)$$

where w_{ji} and b_i are the weights and bias, respectively, C_j is the center of the j^{th} Gaussian, and r_j defines its width.

In this case, the model (5) is completely linear with respect to the weights w_{ji} and biases b_i, but the C_j and r_j must be determined nonlinearly. The proposed systematic identification procedure is the following: *(a)* Initial values of the centers and widths are selected to ensure a good covering of the whole solvent accessibility range. *(b)* This first estimate of the centres and widths allows a linear estimation of the weights and biases thanks to a least squares estimator:

$$\hat{\boldsymbol{\theta}}_{wb} = \underset{\boldsymbol{\theta}_{wb}}{ArgMin}(J) \quad , \text{ with}$$

$$J = \frac{1}{2N}\sum_{s=1}^{N}\left(\Delta\Delta G_{M,s} - \Delta\Delta G_{P,s}\right)^2 \quad , \quad (6)$$

where J is the cost function, θ_{wb} is a vector containing all parameters w_{ji} and b_i (with i=1,2,3 and j=1,2,…,27), N is the number of mutants, $\Delta\Delta G_{M,s}$ is the experimentally measured stability change of mutant s and $\Delta\Delta G_{P,s}$ its predicted value, which is a function of θ_{wb} through (4) and (5). *(c)* Once a first estimation of the whole set of parameters is available, the coefficients are released in order to refine the model. Actually, only the centres and widths are re-estimated while the weights and biases are deduced linearly, with (6), from the new Gaussian parameter values:

$$\hat{\boldsymbol{\theta}}_{Cr} = \underset{\boldsymbol{\theta}_{Cr}}{ArgMin}(J). \quad (7)$$

Note that constraints on the C_j and r_j values are introduced in the estimator to limit the recovery of the Gaussians, in order to avoid redundancy or compensation effects. Different values of the maximal recovery were tested.

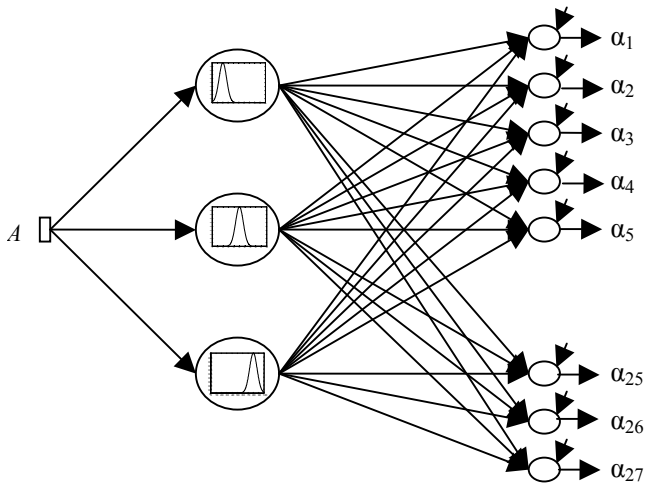

Fig. 1. Radial Basis Function network structure

After these three identification steps, 114 parameters (3 C_j, 3 r_j, 81 w_{ji}, and 27 b_i) have been estimated. As this high parameter amount may lead to overfitting and weak cross validation results, network pruning techniques have to be processed. Classic reduction parameter is preferred as it may be directly interpreted from a biological point of view. For instance, a weight cancellation would reflect the absence of solvent-accessibility influence on the contribution of a potential. Since centres and widths cannot be eliminated without major degradation of the model, only weights and biases are potentially cancellable. As they are estimated in a last linear identification step independently of the nonlinear estimation of the centres and widths, we compute only the covariance on the weights and biases assuming the centres

and widths are well known. The proposed parameter reduction procedure works in numerous steps in order to avoid too fast cancellation of significant parameters. Table 2 gives dimensionless thresholds determined by the trial and error method. Once hardly assessable parameters are cancelled out according to the mentioned thresholds, the remaining parameters are re-identified according to (6-7).

Table 2. Steps of the RBF parameter reduction procedure

Step	1	2	3	4	5	6	7	8
Variance threshold	100	10	5	2.5	1.5	1	1	1
Covariance threshold	100	10	5	2.5	1.5	1	1	1

3.2 MultiLayer Perceptron (MLP)

As the proportionality functions $\alpha_i(A)$ should reproduce the dominating influence of different interactions at the protein surface and in the core, we assume *a priori* that these functions present a sigmoid profile. However, the RBF structure does not ensure such a profile. This is why a second network architecture is considered. It consists in 27 independent perceptrons with one sigmoid hidden neuron and a linear output (Fig. 2). The mathematical expression of this network is

$$\alpha_i(A) = w_i \frac{1}{1+e^{-r_i(A-c_i)}} + b_i, \quad (8)$$

where c_i is the inflexion point of the i^{th} sigmoid, and r_i its slope.

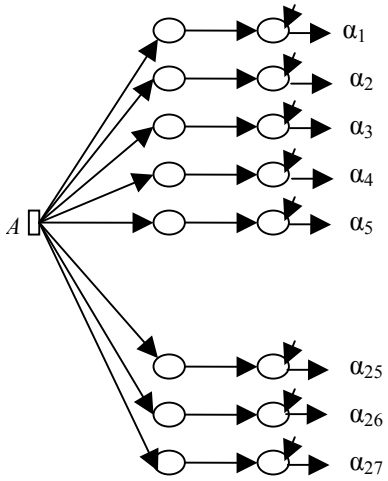

Fig. 2. MultiLayer Perceptron network structure

Regarding the parameter identification, this model is also linear with respect to the weights w_i and biases b_i. Hence, these parameters can be deduced linearly once slopes and inflection points are determined. The identification procedure is similar to that described above for the RBF model: *(a)* The inflection points c_i are initially fixed to 50%, in the middle of the solvent accessibility range. As for the initial slopes r_i, three different values are tested separately: 0.1, 1, and 2.5. *(b)* The weights w_i and biases b_i are estimated linearly according to (6). *(c)* All parameters are released in order to refine the model through non-linear optimisation (7). As for the RBF model, only the hidden parameters are re-estimated while the weights and biases are deduced linearly from the new sigmoid parameter values.

After this identification, 108 parameters (27 c_j, 27 r_j, 27 w_i, and 27 b_i) have been estimated. Parameter reduction is conducted in several steps, as for the RBF model (Table 2). However, contrary to the RBF model, all parameters are considered: hardly assessable slopes and inflection points are set to their initial values, and hardly assessable weights and biases are cancelled out.

4. RESULTS AND DISCUSSION

4.1 Training and validation sets

A set of experimentally characterized mutants was extracted from the ProTherm database (Bava *et al.* 2004). These data were filtered in order to eliminate bad quality inputs, according to the following criteria : *(a)* When the same mutant appears several times in the database, only one value of $\Delta\Delta G_M$ is selected. *(b)* Measurements performed at a pH lower than 6 or higher than 8 are not considered. *(c)* Multiple mutations, mutations involving a proline, and mutations with $\Delta\Delta G_M > 5$ kcal/mol are not considered, as they are likely to induce structural modifications, which are not modelled by PoPMuSiC. *(d)* We consider only mutations whose stability changes are measured with respect to the same reference state, *i.e.* the unfolded conformation. *(e)* The experimental protein structure must be available. *(f)* If the protein is multimeric, free energy changes refer either to the whole protein or to a monomer only. Mutations for which this is not specified are eliminated.

As a result, 555 experimentally characterized mutants are selected. Among these mutants, 409 are chosen randomly to constitute an identification set, with the constraint that they must cover the whole solvent accessibility range. The 146 remaining mutants define the validation set. To ensure that no bias arises from the random selection, four pairs of identification/validation sets are built and used in parallel. All results given below are averages over these four pairs of sets.

4.2 Prediction of stability changes upon mutation

To assess the predictive capabilities of both models, we compare them on the basis of J, the average squared difference between predicted and measured free energy changes $\Delta\Delta G$, which is the cost function of our optimisation procedure (6), and on the linear correlation coefficient between the measured and predicted stability changes. It appears that the MLP model performs slightly better than the RBF model in direct validation, but is somewhat less efficient in cross validation (Table 3).

In Fig. 3, the values of the stability changes predicted with the MLP model are plotted against the corresponding experimentally measured values, for all mutants included in the identification and validation sets, respectively.

Intriguingly, some mutants are systematically poorly modelled, in all pairs of identification/validation sets, and with all values chosen for the maximal Gaussian recovery (in the RBF model) or for the initial slope of the sigmoids (in the MLP model). This does not necessarily imply that the energetic functions and/or the structure of the models are defective. Indeed, the stability changes measured for some mutants may be spoiled by a high experimental error. On the other hand, some mutants may involve a significant modification of the protein's backbone structure, which is not taken into account by our models. Therefore, we propose to exclude these outliers from our datasets and re-identify the models. In practice, for each type of model (RBF and MLP) all mutants showing a deviation larger than 1.5 kcal/mol between the predicted and measured free energy change, in more than 90% of the performed identifications (with different databases, different maximal Gaussian recoveries or different initial slopes of the sigmoids), are rejected.

Table 3. Performances of the RBF and MLP models

	Direct validation		*Cross validation*	
	J [1]	r [2]	J [1]	r [2]
RBF	0.51	0.70	0.49	0.66
MLP	0.49	0.73	0.62	0.59
RBF (without outliers)	0.30	0.74	0.32	0.70
MLP (without outliers)	0.33	0.79	0.39	0.74

[1] J is the value of the cost function (6).
[2] r is the correlation coefficient between $\Delta\Delta G_M$ and $\Delta\Delta G_P$ on all mutants of the identification or validation sets.

As expected, after rejection of 81 mutants for the RBF model, or 50 mutants for the MLP model, the performances are notably enhanced (Table 3): the cost functions values are lower and the correlation coefficients higher. We also observe that, relatively to the RBF model, the cross validation performances of the MLP model are improved. Indeed, both models appear mostly equivalent, with slightly better values of the cost function for the RBF model, and slightly better correlation coefficients for the MLP model.

4.3 Biophysical interpretation of the model

A significant advantage of the MLP model over the RBF model is that the proportionality coefficients $\alpha_i(A)$ evolve monotonously with A, which allows an easier interpretation in terms of their biophysical significance.

A few examples of the dependence of α_i on A are given in Fig. 4. $\Delta\Delta W_{st}$ describes interactions between amino acids close to each other in the sequence. The weight of these interactions appears larger on the surface than in the core of the proteins. On the other hand, the weight of $\Delta\Delta W_{sd}$, which is dominated by the hydrophobic effect, is larger in the core. These trends are in agreement with a previous study (Gilis and Rooman, 2000). Furthermore, as expected, the α_i corresponding to $\Delta\Delta W_{sds}$, which describes specific interactions between amino acids close to each other in space, remains constant. The α_i profile corresponding to ΔV_1 is intriguing, for steric stresses should be more destabilizing in the core, and α_i should thus be decreasing. As for ΔV_2, the creation of a cavity in the core of a protein is unfavourable (Eriksonn *et al*, 1992), which is correctly modelled by the increase of the corresponding α_i. Finally, the independent term is positive at small values of A, and tends to zero when A increases, which indicates that mutations in the core are usually more destabilizing than expected. Note that this term and ΔV_1 might compensate each other, which could explain the somewhat surprising behaviour of ΔV_1.

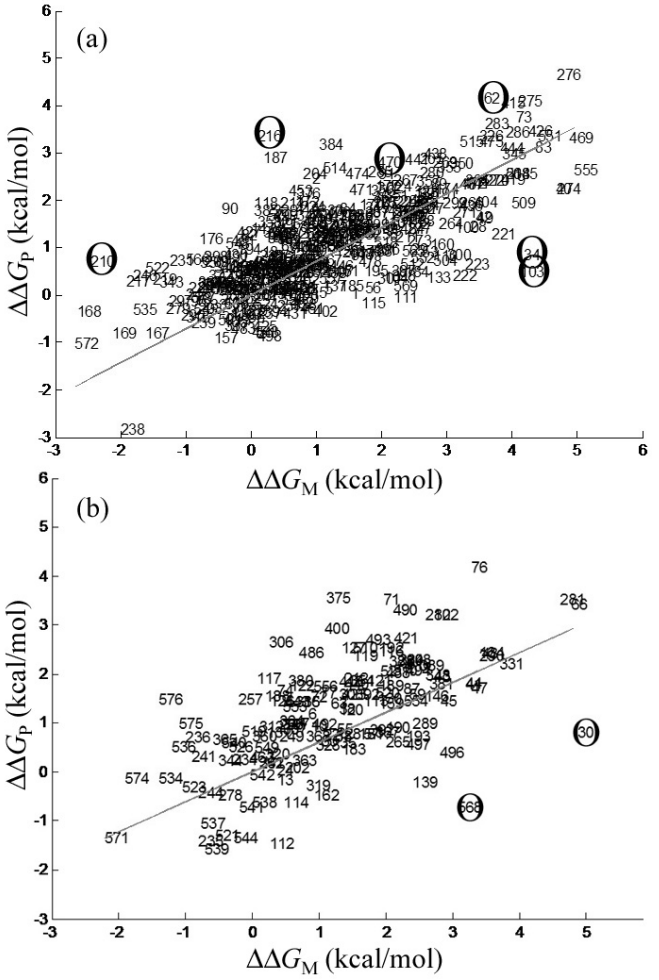

Fig. 3. Correlation between the measured values of $\Delta\Delta G$, and those predicted by the MLP model, in (a) the identification set, and (b) the validation set. Encircled points are systematically poorly modelled.

CONCLUSIONS

The combination of new energetic functions with neural network models that provide weighting coefficients as functions of the solvent accessibility allowed us to devise a new version of the PoPMuSiC program. Although the correlation between predicted and measured stability changes is slightly lower than in the previous tests, it must be noted that the performances of the first version of PoPMuSiC are significantly lower on the new dataset of mutants presented

here. The comparison of the RBF and MLP models results in favor of the MLP model, as only 50 outliers must be rejected, against 81 for the RBF model, to achieve a similar predictive power. In addition, the biophysical interpretation of the MLP model is much more straightforward.

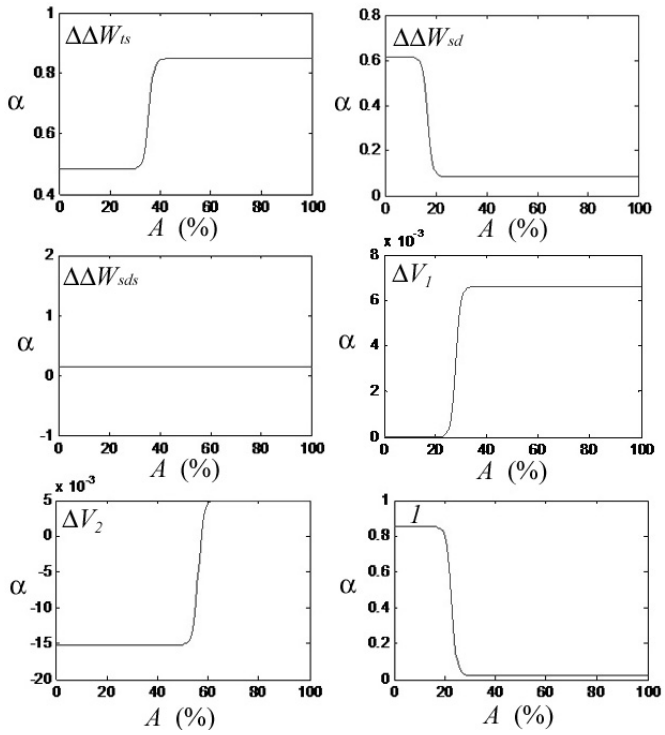

Fig. 4. Evolution of the proportionality coefficients α_i (4) of the MLP model, as a function of the solvent accessibility A. The energetic function to which each α_i refers is indicated.

ACKNOWLEGMENTS

We acknowledge support from the Belgian State Science Policy Office through an Interuniversity Attraction Poles Programme (DYSCO), from the Belgian Fund for Scientific Research (FRS) through an FRFC project, and from the BioXpr company. YD benefits from a First-Postdoc grant of the Walloon region (PROMeTHe) and MR is Research Director at the FRS.

REFERENCES

Basch, P.A., U.C. Singh, R. Langridge and P.A. Kollman (1987). Free energy calculations by computer simulation. *Science*, **236**, 564–568.

Bava, K.A., M.M. Gromiha, H. Uedaira, K. Kitajima and A. Sarai (2004). ProTherm, version 4.0: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res.* **32**, D120-D121.

Capriotti, E., P. Fariselli and R. Casadio (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, **33**, W306-W310.

Creighton, T.E. (1993). *Proteins: structures and molecular properties,* W.H. Freeman and Company, New York.

Dehouck, Y., D. Gilis and M. Rooman (2006). A new generation of statistical potentials for proteins, *Biophysical Journal*, **90**, 4010-4017.

Eriksson, A.E., W.A. Baase, X.-J. Zhang, D.W. Heinz, M. Blaber, E.P. Baldwin and B.W. Matthews (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.

Gilis, D., and M. Rooman (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins. *Protein Engineering*, **13**, 849-856.

Guerois, R., J.E. Nielsen and L. Serrano (2002). Predicting changes in the stability of proteins and protein complexes : a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369-387.

Huang, L.-T., M.M. Gromiha and S.-Y. Ho (2007). Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model, *J. Mol. Model.*, **13**, 879-890.

Kabsch, W. and C. Sander (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.

Kwasigroch, J.M., D. Gilis, Y. Dehouck and M. Rooman (2002) PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics*. **18**,1701-1702.

Lehninger, A.L., D.L. Nelson and M.M. Cox (1993). *Principles of Biochemistry, 2nd edition*, Worth publishers, New York.

Miyazawa, S. and L. Jernigan (1994). Protein stability for single substitution mutants and the extent of local compactness in the denatured state. *Protein Eng*. **7**, 1209–1220.

Muñoz, V. and L. Serrano (1994). Intrinsic secondary structure propensities of the amino acids, using statistical f-y matrices: comparison with experimental data. *Proteins: Struct. Funct. Genet*., **20**, 301–311.

Parthiban, V., M.M. Gromiha and D. Schomburg (2006). CUPSAT:prediction of protein stability upon point mutations. *Nucleic Acids Research*, **34**, W239-W242.

Rooman, M.J., J.P. Kocher and S.J. Wodak (1991). Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions, *J. Mol. Biol*., **221**, 961-979.

Rose, G.D., A.R. Geselowitz, G.J. Lesser, R.H. Lee and M.H. Zehfus (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, **29**, 834-838.

Shortle, D., W.E. Stites and A.K. Meeker (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**, 8033–8041.

Sippl, M.J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**, 229–235.

Tidor, B. and M. Karplus (1991). Simulation analysis of the stability mutant R96H of T4 lysozyme. *Biochemistry*, **30**, 3217–3228.

van Gunsteren, W.F. and A.E. Mark (1992). Prediction of the activity and stability effects of site-directed mutagenesis on a protein core. *J. Mol. Biol*., **227**, 389–395.