

# PREDICTING FREE ENERGY CONTRIBUTIONS TO THE CONFORMATIONAL STABILITY OF FOLDED PROTEINS FROM THE RESIDUE SEQUENCE WITH RADIAL BASIS FUNCTION NETWORKS

Rita Casadio, Mario Compiani<sup>o</sup>, Piero Fariselli and Francesco Vivarelli

Laboratory of Biophysics, Dept. of Biology, University of Bologna, Bologna, Italy, and <sup>o</sup> Dept. of Chemistry, University of Camerino, Camerino, Italy.  
E-mail:G4XBO3B1@CINE88.CINECA.IT

## Abstract

Radial basis function neural networks are trained on a data base comprising 38 globular proteins of well resolved crystallographic structure and the corresponding free energy contributions to the overall protein stability (as computed partially from chrystallographic analysis and partially with multiple regression from experimental thermodynamic data by Ponnuswamy and Gromiha (1994)). Starting from the residue sequence and using as input code the percentage of each residue and the total residue number of the protein, it is found with a cross-validation method that neural networks can optimally predict the free energy contributions due to hydrogen bonds, hydrophobic interactions and the unfolded state. Terms due to electrostatic and disulfide bonding free energies are poorly predicted. This is so also when other input codes, including the percentage of secondary structure type of the protein and/or residue-pair information are used. Furthermore, trained on the computed and/or experimental  $\Delta G$  values of the data base, neural networks predict a conformational stability ranging from about 10 to 20 kcal mol<sup>-1</sup> rather independently of the residue sequence, with an average error per protein of about 9 kcal mol<sup>-1</sup>.

## Introduction

The aim of "structural" thermodynamics is to establish the relationship between structural information contained in the data base of protein structures and the thermodynamic stability of the folded protein (Freire 1993). This will provide the necessary theoretical background for rationalizing protein engineering (Mathews 1993). However the number of well resolved structures in the data base (presently about 300) largely exceeds the number of proteins whose thermodynamic stability has been experimentally assessed (14 proteins at 298 K, Privalov and Gill 1988; Livingstone et al., 1991). A more substantial gap exists between the number of primary sequences presently available and the number of crystallized proteins. Conformational stability of proteins results from the competition between opposite dominant

contributions, namely folding and unfolding free energies (Dill, 1990), as reflected also by the temperature dependence of the  $\Delta G$  stability values (Murphy et al., 1990). For this reason, it is generally believed that proteins are only marginally stable at room temperature (Privalov and Gill, 1988; Murphy et al., 1990) with a typical  $\Delta G$  value in the range of 5 - 20 kcal mol<sup>-1</sup> of protein. Contributions to conformational stability are mainly due to hydrogen bonding, hydrophobicity, solvation effects, ion pairs, van der Waals interactions and disulphide bonds. Each contribution can be evaluated from the protein structure or the chemico-physical properties of its residues (Dill, 1990). Particularly the hydrophobic effect has been related to the difference in hydration of polar and non polar groups upon unfolding (Livingstone et al., 1991; Makhatadze and Privalov, 1993; Privalov and Makhatadze, 1993). The free energy cost of burying polar groups in the protein interior nearly compensates the stabilizing contribution of the hydrophobic effect and could account for the marginal stability of proteins (Yang et al., 1992). Recently all the different contributions to the conformational stability of a set of 38 well resolved proteins, comprising the 14 proteins whose thermodynamic data are available, have been determined from the crystal structures and with linear multiple regression from the experimental data (Ponnuswamy and Gromiha, 1994). We use these data to investigate with radial basis function (RBF) networks whether and to which extent the residue sequence of the protein is related to the stability of the folded state. RBF networks are particularly suited in finding smooth representation of underlying trends in data sets of relatively small size (Bishop, 1995). Our results indicate that only the free energy contributions due to hydrophobic effects, hydrogen bonding and unfolded state can be accurately determined from the residue sequence of the protein.

## The Data Base of Proteins and their Free Energy Values

The data base of globular proteins comprises 38 proteins known with an atomic resolution  $\leq 2.5 \text{ \AA}$  (labelled with the Brookhaven code). The set includes the 14 proteins, whose thermodynamic data are available (Privalov and Gill, 1988; Livingstone et al., 1991). The primary structure content is taken from the corresponding Protein Data Bank files of Brookhaven and the percentage of secondary structure is evaluated with the Define Secondary Structure of Proteins (DSSP) program of Kabsch and Sander (1983).

The data base of free energy values is from Ponnuswamy and Gromiha (1994). These data for each of the 38 proteins have been either evaluated from the corresponding crystal structures or empirically computed by extrapolating from the available thermodynamic data. More specifically, each contribution was evaluated as summarized in the following.

The conformational stability of the native state of a globular protein is defined and calculated as:

$$\Delta G = G_f - G_u \quad (1)$$

where  $G_f$  and  $G_u$  are the free energies of the folded and unfolded state of the protein.

$G_u$  was computed summing up the entropic contribution suggested by Tanford (1980) ( $1.2 \text{ kcal mol}^{-1}$  per residue) and a non entropic term, due to weak hydrogen bonding interaction between chain and solvent molecules. This was assumed to be half the contribution of the folded state ( $G_{hb}$ ), according to:

$$G_u = \left( 1.2 \cdot N + \frac{1}{2} \cdot G_{hb} \right) \quad (2)$$

where  $N$  is the number of residues in the protein.

The free energy of folding is in turn evaluated as the total sum of different contributions:

$$G_f = G_{hy} + G_{hb} + G_{el} + G_{ss} + G_{vw} \quad (3)$$

where the subscripts identify, respectively, hydrophobic (hy), hydrogen bonding (hb), electrostatic (el), disulphide bonding (ss) and van der Waals (vw) free energies.

The hydrophobic free energy of protein folding was expressed as

$$G_{hy} = \sum_i \Delta\sigma_i \left[ A_i(\text{folded}) - A_i(\text{unfolded}) \right] \quad (4)$$

where  $\Delta\sigma_i$  is an atomic solvation parameter evaluated with the method of Eisenberg and McLachlan (1986).  $A_i(\text{folded})$  and  $A_i(\text{unfolded})$  represent respectively the accessible areas of each atom in the folded and unfolded state of the protein and were evaluated using the ACCESS program of Richmond and Richards (1978).

The free energy from hydrogen bonds was computed considering an approximate value of  $1 \text{ kcal mol}^{-1}$  for each

potential hydrogen bond (Ben - Naim, 1991) and considering that the actual number of hydrogen bonds in a protein ( $N_{hb}$ ) is the number of hydrogen bonds minus the number of ion pairs in the protein (whose contribution is included in  $G_{el}$ , in Eq.6) (Privalov and Gill, 1988):

$$G_{hb} = 1 \cdot N_{hb} \quad (5)$$

The electrostatic free energy ( $G_{el}$ ) was taken to be the sum of three terms due to the number of ion-pairs and to the number ( $N_{ch}$ ) of charge-helix dipole interactions (about  $1.6 \text{ kcal mol}^{-1}$ ). The ion pairs were grouped in surface ( $N_{si}$ ) and buried ( $N_{bi}$ ) (contributing  $1 \text{ kcal mol}^{-1}$  and  $3 \text{ kcal mol}^{-1}$ , respectively)

$$G_{el} = 3 \cdot N_{bi} + 1 \cdot N_{si} + 1.6 \cdot N_{ch} \quad (6)$$

The contribution from disulphide bridges was taken proportional to the number of disulphide bonds ( $N_{ss}$ ) (with a value of  $2.3 \text{ kcal mol}^{-1}$  per single bond (Thornton, 1981)):

$$G_{ss} = 2.3 \cdot N_{ss} \quad (7)$$

$G_{vw}$  in Eq. 3 was evaluated as the complement of all the free energy contributions listed above (Eqs. 4 - 7) to the available thermodynamic data ( $\Delta G$ ) of 14 proteins (Privalov and Gill, 1988).  $G_{vw}$  is found to be linearly correlated with the number of protein residues. A multiple regression analysis was performed to assess the relative weight of each contribution to the conformational stability of the folded state.  $\Delta G$  was predicted to range from about 4 to  $40 \text{ kcal mol}^{-1}$ , with an excellent agreement between the experimental and theoretical values for each of the 14 proteins (when they were not included in the input of the regression analysis).

## The Architecture of Radial Basis Function (RBF) Neural Networks

RBF neural networks are based on the notion that an arbitrary function  $y(x)$  can be approximated as the linear superposition of a set of localized basis functions  $\Phi_j(x)$ , one for each data point in the training set. The network structure consists of a set of  $m$  hidden nodes connecting the  $n$  input vectors to the output nodes in the output layer; the output of the  $k^{\text{th}}$  neuron is given by

$$y^k(x) = \sum_{j=1}^m w_{kj} \cdot \Phi_j(x) \quad (8)$$

$\Phi_j(x)$  is a radially symmetric function, which represents the activation of hidden unit  $j$  when the network is presented with input vector  $x$ . The basis function is chosen to be a Gaussian

$$\Phi_j(x) = \exp\left(-\frac{|x - \mu_j|^2}{2 \cdot \sigma_j^2}\right) \quad (9)$$

where  $\sigma_j$  is the width parameter and  $\mu_j$  is a vector representing the center of the  $j^{\text{th}}$  basis function.

In this architecture the adjustable parameters are the centers  $\mu_j$  and widths  $\sigma_j$  of the Gaussian functions and the weights  $w_{kj}$  in Eq.8. The width parameters are initially set equal to the average distance between the basis function centers. The  $\mu_j$  vectors, in turn, are selected with an iterative approach based on the K-means algorithm, grouping the input patterns into clusters with one basis function center acting as the representative vector for each cluster. The weights are evaluated by minimizing the deviation from the desired output  $t_k^q$  for the  $q^{\text{th}}$  input pattern

$$E = \frac{1}{n} \cdot \sum_q \sum_k \left( y_q^k(x) - t_q^k \right)^2 \quad (10)$$

where  $n$  is the number of patterns.

The weights are set to their optimal values once the basis function parameters have been determined.

The most efficient networks in generalization (with optimal number of hidden nodes) are chosen after an exhaustive search in the space of architectures (see Table 1).

Different inputs to the networks are considered: code 1, percentage of each residue in the protein sequence and numbers of residues; code 2, percentage of each residue in the protein sequence and percentage of secondary structure ( $\alpha$ -helix,  $\beta$ -sheet and coil) for each protein; code 3, a matrix input code describing contiguous residue pairs in the sequence; code 4, code 3 supplemented with the percentage of secondary structure.

The 38 proteins of the training set are divided into 8 subsets, 7 of which contain 5 proteins and one three proteins. Alternatively one subset is excluded from the training procedure and used to test the predictive performance of the network (cross-validation). The errors in generalization (in Table 1) are the average of the square root of the errors (Eq.10) for the different subsets.

**Table 1.** Performance of optimized RBF networks in predicting the free energy terms using different input codings and architectures.

Data Base	Input coding								
	Code 1	Code 2	Code 3	Code 4	Code 1	Code 2	Code 3	Code 4	
Average values	Hidden nodes	Error	Hidden nodes	Error	Hidden nodes	Error	Hidden nodes	Error	
$G_u$	264 ± 152	9	5.5	10	24.7	4	26.2	10	24.7
$G_{hy}$	115 ± 71	13	11.7	10	14.9	4	13.8	10	14.9
$G_{hb}$	116 ± 68	4	5.4	10	11.5	4	10.4	10	11.5
$G_{el}$	14 ± 9	2	5.5	2	5.9	6	5.9	2	5.5
$G_{ss}$	3 ± 4	2	3.7	2	3.0	6	3.0	2	3.7
$G_{vw}$	33 ± 14	4	0.3	9	1.9	4	1.8	9	1.9
$\Delta G$	17 ± 10	2	9.5	2	9.4	2	9.3	2	9.5

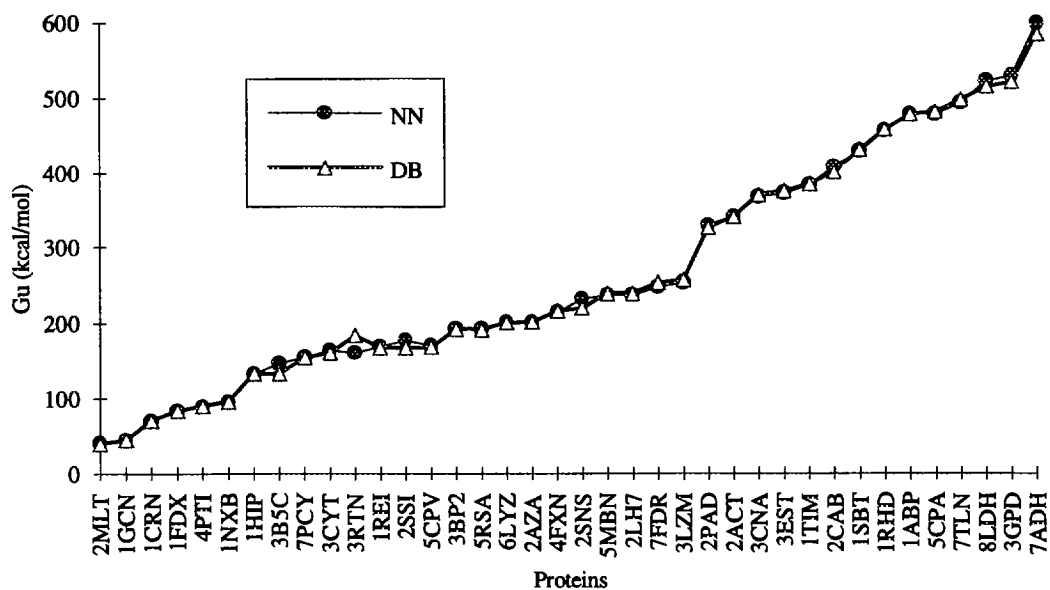
Code 1 = percentage of residues and length of the protein.

Code 2 = Code 1 + percentage of secondary structures ( $\alpha$ -helix,  $\beta$ -sheet and random-coil).

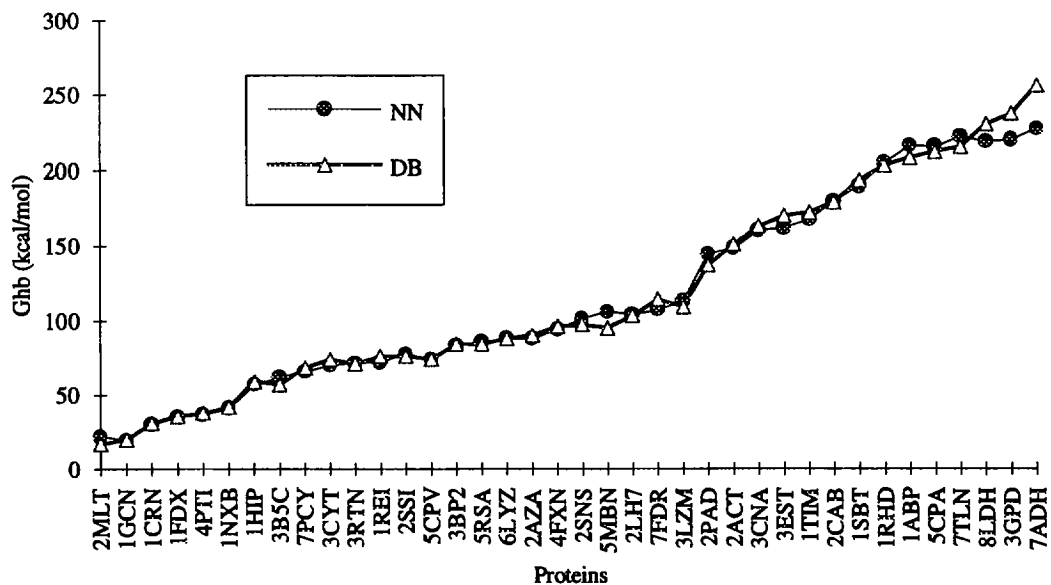
Code 3 = Residue pair matrix (20x20).

Code 4 = Code 3 + percentage of secondary structures.

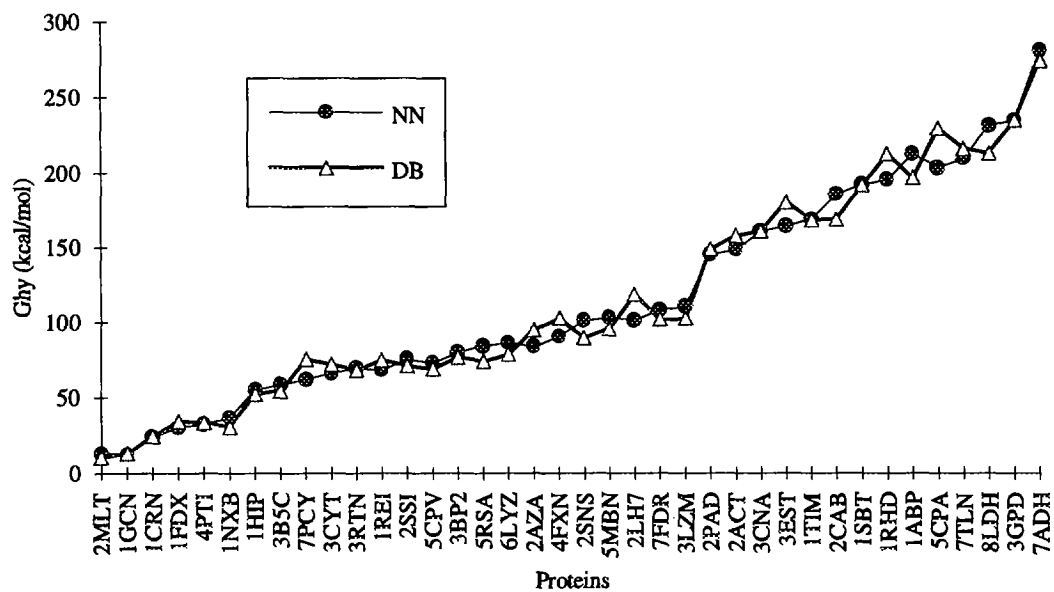
Errors and average free energy values of the Data base (± standard deviation) are in kcal/mol. The shaded areas indicate the most efficient RBF network architectures used to draw the plots in Figs. 1-7.



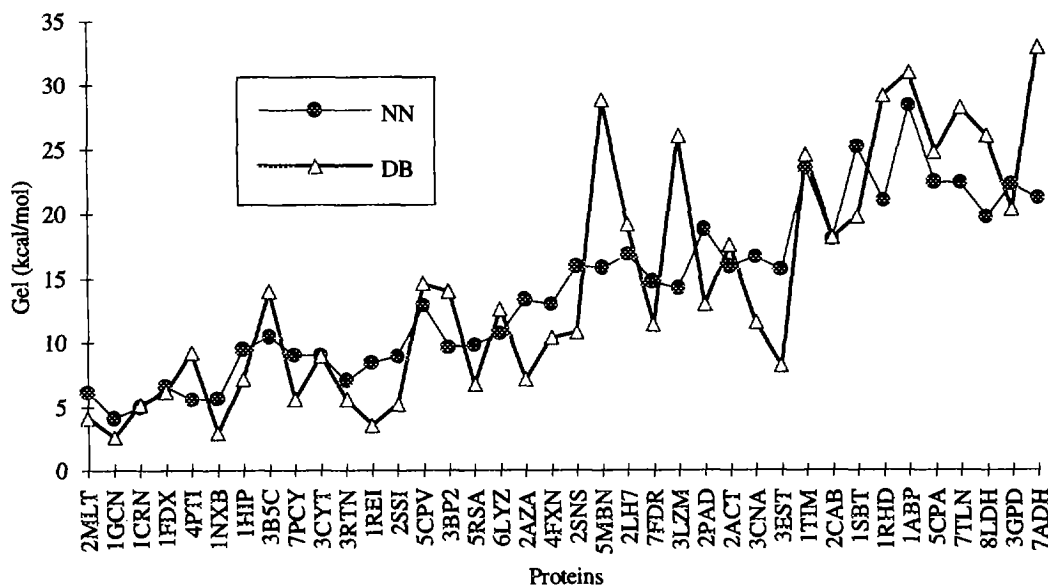
**Fig. 1** Free energy values of the unfolded state ( $G_u$ ); DB (Data base) = target values; NN = neural network predictions; proteins are ordered according to increasing number of residues.



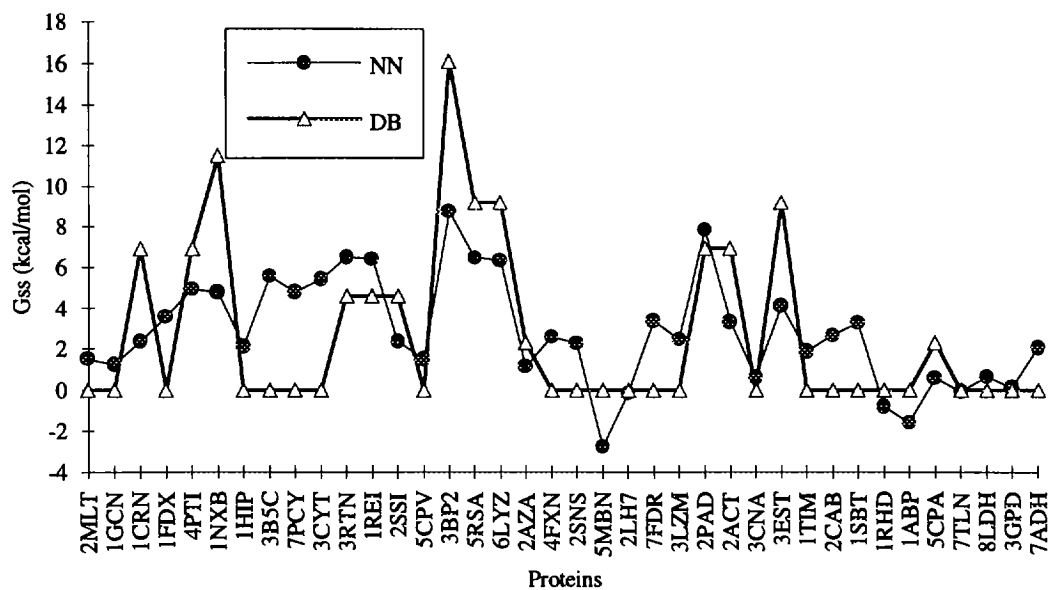
**Fig. 2** Free energy values of hydrogen bonding ( $G_{hb}$ ).



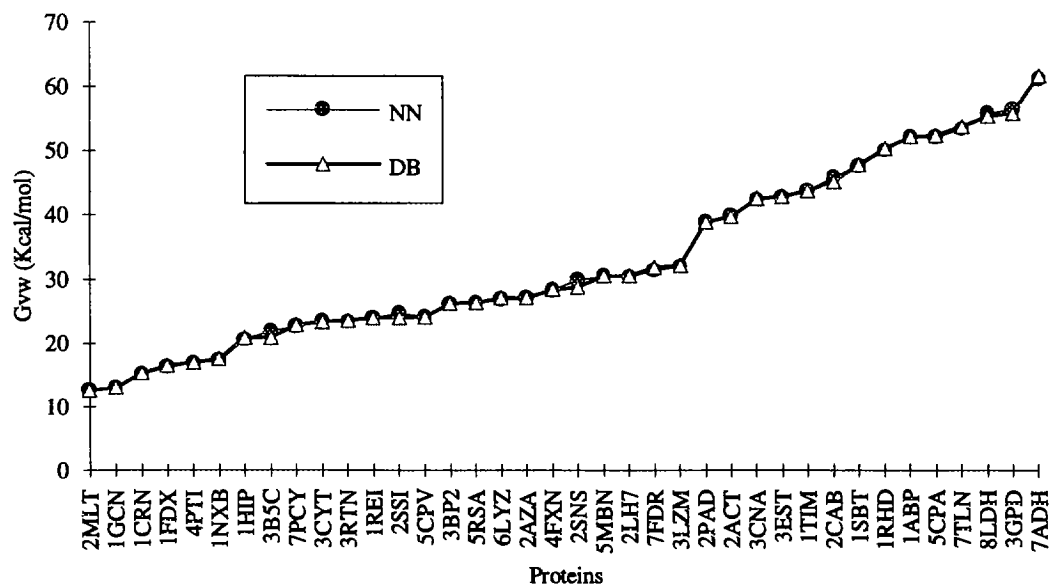
**Fig. 3** Hydrophobic free energy values ( $G_{hy}$ ).



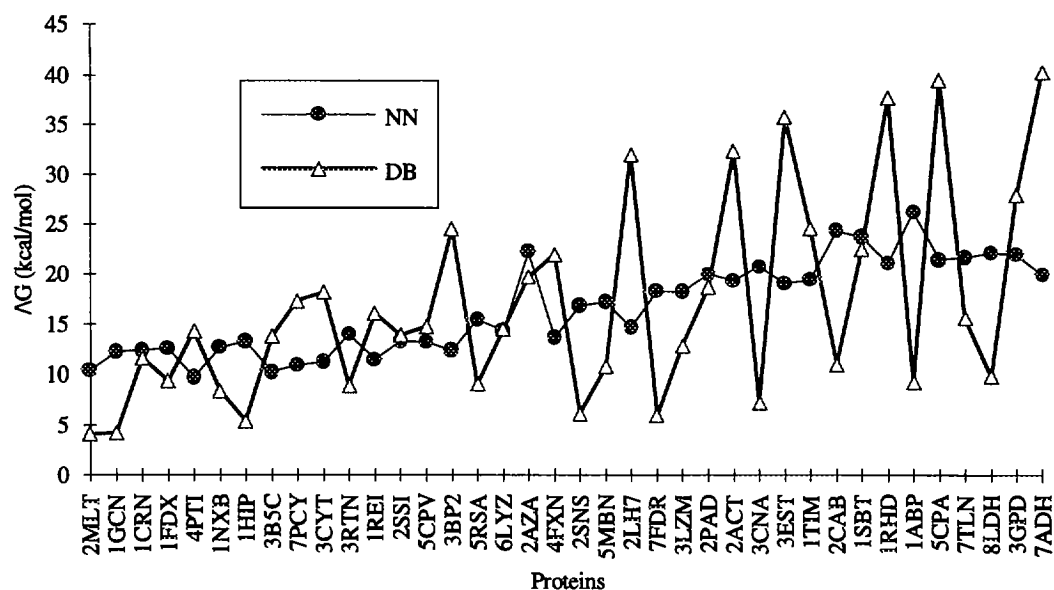
**Fig. 4** Electrostatic free energy values ( $G_{el}$ ).



**Fig. 5** Disulphide bonding free energy values ( $G_{ss}$ ).



**Fig. 6** Van der Waals free energy values ( $G_{vw}$ ).



**Fig. 7** Conformational stability values ( $\Delta G$ ).

## Results and Discussion

Using the input code 1 based on the primary sequence of the protein RBF neural networks can well predict the free energy contributions due to the unfolded state, hydrogen bonds and hydrophobic interactions (Eqs. 2, 4 and 5, respectively). The results (shown in Figs. 1 - 3) indicate that this input code contains sufficient information on the difference between solvent accessible surface area of the folded and unfolded state (Eq. 4) and on the number of actual hydrogen bonds in the protein (Eqs. 2 and 5). RBF neural networks learn rules of association from the input to the free energy and extrapolate values for the never seen before proteins in the testing set. The accuracy of the prediction is indicated by the low mean error per protein compared to the average free energy value of the data set (Table 1).  $G_u$ ,  $G_{hy}$  and  $G_{hb}$  are functions of the type and number of residues (Eqs. 2, 4 and 5) and increase with increasing size of the protein. This trend is also correctly predicted by the network (the proteins are ordered with increasing residue number in all the figures).

On the other hand, contributions arising from electrostatic free energy and disulphide bonding, are poorly predicted using the same input code (Figs. 4 - 5 and Table 1). Depending on the number of ion pairs (buried or unburied), charge/helix dipole interactions, and disulphide bonds, these free energy terms are highly specific for each protein. None of these properties correlate with the protein size and number and/or type of residue (Eqs. 6 and 7). Networks trained with a more extended input code

including residue-pairs information (code 3) and the composition in secondary structure (code 4), tend to perform on  $G_{el}$  and  $G_{ss}$  slightly better improving the correlation between the data base (DB) and the predicted (NN) signals (Figs. 4 and 5). The mean error per protein is however remarkably higher than for the other contributions (Table 1).

Considering the empirical free energy values ( $G_{vw}$  and  $\Delta G$ ) contained in the data base, networks well predict  $G_{vw}$ , which was found to be a linear function of the number of protein residues (Ponnuswamy and Gromiha, 1994) (Table 1 and Fig. 6). The prediction of the conformational stability value (which is not a linear function of the residue type and numbers) is poor and is only slightly improved when the input codes including residue pairs information and/or the percentage of secondary structure are used (Fig 7 and Table 1). Notably the network follows the average trend but with a minimal correlation with the data base (Fig. 7). A similar lack of correlation obtains when the training is performed on a set comprising only the 14 proteins whose  $\Delta G$  was experimentally determined (data not shown). The resulting correlation in Fig. 7 is smaller than that expected taking into account the errors made in predicting  $G_{el}$  and  $G_{ss}$ .

Consistently with previous suggestions (Dill, 1990; Yang et al., 1992), our analysis with neural networks indicates that  $\Delta G$  values are likely to be found in the range of 10 - 20 kcal.mol<sup>-1</sup>, rather independently of the protein size and content of residues and/or secondary structure types.

## References

- Ben - Naim, A. 1991. The role of hydrogen bonds in protein folding and protein association. *J Phys Chem* 95:1437 - 1444
- Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford
- Dill, K. A. 1990. Dominant forces in protein folding. *Biochemistry* 29, 7133 - 7155
- Eisenberg, D.; and McLachlan, A. D. 1986. Solvation energy in protein folding and binding. *Nature* 319: 189 - 203
- Freire, E. 1993. Structural Thermodynamics: prediction of protein stability and protein binding energy. *Arch. Biochem. Biophys.* 303: 181-184.
- Kabsch, W.; and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded geometrical feature. *Biopolymers* 22: 2577 - 2637
- Livingstone, J. R.; Spolar, R. S.; and Record, T. M. 1991. Contribution to the thermodynamics of protein folding from the reduction in water-accessible nonpolar surface area. *Biochemistry* 30: 4237 - 4244
- Mathews, B. W. 1993. Structural and genetical analysis of protein stability. *Annu Rev Biochem* 62: 139 - 160
- Murphy, K. P.; Privalov, P. L.; and Gill, S. J. 1990. Common features of protein unfolding and the dissolution of hydrophobic compounds. *Science* 247: 559 - 561
- Ponnuswamy, P. K.; and Gromiha, M. M. 1994. On the conformational stability of folded proteins. *J Theor Biol* 166: 63 - 74
- Privalov, P. L.; and Gill, S. J. (1988) Stability of protein structure and hydrophobic interaction. *Adv Prot Chem* 39: 191 - 234
- Privalov, P. L.; and Makhatadze, G. I. 1990. Heat capacity of proteins. II Partial molar heat capacity of the unfolded polypeptide chain of proteins: protein unfolding effects. *J Mol Biol* 213: 385 - 391
- Richmond, T. J.; and Richards, F. M. 1978. Packing of alpha-helices: geometrical constraints and contact areas. *J Mol Biol* 119: 537 - 555
- Thornton, J. M. 1981. Disulphide bridges in globular proteins. *J Mol Biol* 151: 261 - 287
- Tanford, C. 1980. *The Hydrophobic Effect*. John Wiley, New York
- Yang, A. S.; Sharp, K. A.; and Honig, B. 1992. Analysis of the heat capacity dependence of protein folding. *J Mol Biol* 227: 889 - 900