BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii54–ii58
doi:10.1093/bioinformatics/bti1109

Protein Structure and Function

Predicting protein stability changes from sequences using support vector machines

Emidio Capriotti, Piero Fariselli, Remo Calabrese and Rita Casadio*

Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy

ABSTRACT

Motivation: The prediction of protein stability change upon mutations is key to understanding protein folding and misfolding. At present, methods are available to predict stability changes only when the atomic structure of the protein is available. Methods addressing the same task starting from the protein sequence are, however, necessary in order to complete genome annotation, especially in relation to single nucleotide polymorphisms (SNPs) and related diseases.

Results: We develop a method based on support vector machines that, starting from the protein sequence, predicts the sign and the value of free energy stability change upon single point mutation. We show that the accuracy of our predictor is as high as 77% in the specific task of predicting the $\Delta\Delta G$ sign related to the corresponding protein stability. When predicting the $\Delta\Delta G$ values, a satisfactory correlation agreement with the experimental data is also found. As a final blind benchmark, the predictor is applied to proteins with a set of disease-related SNPs, for which thermodynamic data are also known. We found that our predictions corroborate the view that disease-related mutations correspond to a decrease in protein stability.

Availability: http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi

Contact: casadio@alma.unibo.it

1 INTRODUCTION

Protein stability change upon site-specific mutations is a relevant problem both for protein design and for the comprehension of protein function (Daggett and Fersht, 2003). For this reason, different methods have been described to predict stability changes observed upon residue substitutions in the original protein sequence. They are based mainly on the development of different energy functions and are suited to computing the stability free energy changes in protein structures when mutating one residue at a time in the sequence (Prevost *et al*., 1991; Topham *et al*., 1997; Pitera and Kollman, 2000; Gilis and Rooman, 1997; Kwasigroch *et al*., 2002; Funahashi *et al*., 2001; Guerois *et al*., 2002; Zhou and Zhou, 2002). An alternative approach based on a neural network (NN) system was recently proposed (Capriotti *et al*., 2004). In this application, instead of directly estimating the relative stability changes upon protein mutation (the $\Delta\Delta G$ value), an NN predicts the direction towards which the mutation shifts the stability of the protein (namely the sign of $\Delta\Delta G$). It could be towards a positive or negative $\Delta\Delta G$ value, corresponding to an increase or decrease of stability, respectively. This prediction is sufficient to evaluate the overall effect of the mutation on the protein stability.

*To whom correspondence should be addressed.

Other relevant thermodynamic parameters in mutagenesis are experimental conditions such as pH and temperature (Bava *et al*., 2004). In this respect, energy-based methods need to fit these parameters assuming that the mutations are carried out at physiological conditions. This problem was also overpassed by the machine learning approach (Capriotti *et al*., 2004), which takes these variables as input.

All the methods mentioned above are, however, limited in that prediction can be carried out only when the protein 3D structure is available. For wide-scale genome analysis, it is necessary to develop applications that can predict stability variation upon mutation starting from the protein sequence. This is particularly relevant to assessing whether a given mutation may or may not lead to protein misfolding and diseases (Dobson, 2003).

In this paper we develop a method based on support vector machines (SVMs) that predicts protein stability changes due to single point mutation starting from the sequence. Owing to the availability of a large database of thermodynamic data for mutated proteins (Bava *et al*., 2004) we are able to show that for the specific task of predicting the $\Delta\Delta G$ sign, our method reaches an accuracy value as high as 77% and a satisfactory correlation agreement when assigning the $\Delta\Delta G$ values. Furthermore, we show that the prediction of protein stability decrease correlates well with a blind set of thermodynamic measurements performed with disease-related mutated chains of the prion and transthyretin proteins.

2 METHODS

2.1 The protein database

Our data set is derived from the current release (December 2004) of the Thermodynamic Database for Proteins and Mutants (ProTherm by Bava *et al*., 2004). The data set of proteins was extracted from ProTherm with the following constraints:

(1) the $\Delta\Delta G$ value has been experimentally detected and is reported in the database;

(2) the data are relative to single mutations (no multiple mutations have been taken into account).

After this filtering procedure, we ended up with a data set consisting of 2048 different single mutations obtained from 64 different protein sequences. The final set is available at http://gpcr2.biocomp.unibo.it/∼emidio/I-Mutant2.0/dbMutSeq.html.

2.2 The data set of disease-related mutations

In order to test our predictor on the task of predicting whether diseases induced by single point mutations can destabilize the protein folding, we collected mutations for two experimentally well characterized proteins: the prion protein (PRIO_HUMAN) and transthyretin (TTHY_HUMAN). We collected

all the disease-related mutations known to destabilize the protein folding for which thermodynamic data could also be found in the literature. We included those disease-related mutations that have been reported as having promoted conformational changes and whose 3D structure has been deposited in the Protein Data Bank (PDB). We ended up with 20 mutations for the two proteins, among which some are associated with diseases such as Gerstmann–Strussler and Creutzfeldt–Jakob syndromes and some with amyloidosis for prion and transthyretin, respectively. These data were used as a blind test for our predictor.

2.3 The predictor

We address two different tasks: (1) the prediction of the sign of the protein stability change upon single point mutation and (2) the prediction of the $\Delta\Delta G$ value. The former case is a classification task, discriminating two classes as described before (Capriotti *et al*., 2004). In the latter case we deal with a regression-fitting problem. When developing methods addressing both tasks, we adopted the same type of input. Thus, and for the user, the only difference between tools predicting the $\Delta\Delta G$ sign and those predicting the $\Delta\Delta G$ values is the output type.

Two machine learning algorithms were implemented: (1) a standard feed-forward NN, with the back-propagation algorithm as a learning procedure, and (2) an SVM with several kernels.

For the classification task, the NN architecture consists of a one-layer perceptron with two hidden nodes and one output node that codifies for the increased protein stability ($\Delta\Delta G \geq 0$, desired output set to 1) or for the destabilizing mutation ($\Delta\Delta G < 0$, desired output set to 0). The decision threshold is set equal to 0.5. The same classification labeling and decision threshold are used for the SVMs. Similar to the previous method for predicting stability changes starting from the protein structure (Capriotti *et al*., 2004), the input vectors (the same for NN and SVM) consist of 42 values. The first two input values account for the temperature and the pH at which the stability of the mutated protein was measured. The next 20 (the 20 residue types) explicitly define the 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). The final 20 input values encode the sequence residue environment (again the 20 neurons represent the 20 residue types). Each of these input neurons is provided with the number of the encoded residue type, to be found inside a window centered at the residue that undergoes the mutation and that symmetrically spans the sequence to the left (N-terminus) and to the right (C-terminus) with variable lengths from 7 to 23 residues.

The NNs are our own implemented software. For the SVM implementation we use LIBSVM (http://www.csie.ntu.edu.tw/∼cjlin/). We tested the following available kernels:

Linear $K(x_i, x_j) = x_i T x_j$;
Polynomial $K(x_i, x_j) = (G x_i T x_j + r)^d$;
Sigmoid $K(x_i, x_j) = \tan \mathrm{h}(G x_i T x_j + r)$;
RBF $K(x_i, x_j) = \exp(-G||x_i - x_j||^2)$.

When assigning the $\Delta\Delta G$ values, only the SVM with the RBF kernel is considered. The same input of the classification task is adopted. In this case the SVMs directly compute the regression and the output is the predicted $\Delta\Delta G$ value for a given mutation.

2.4 Scoring the performance

Results obtained with NNs and SVMs are evaluated using a cross-validation procedure on the data set. The reported data for the classification and regression tasks are obtained adopting a 20-fold cross-validation procedure; we also adopted larger and smaller divisions (from 10- to 30-fold cross-validation) in order to asses the stability of the methods and found no difference. Grouping of the data into sets for cross-validation was performed in such a way that the positive and the negative examples respected the original distribution of the whole set. Furthermore, we kept the same mutations (when reported at different experimental conditions) in the same set to prevent an overestimation of the results. For each tested method we adopted the same cross-validation sets; thus, results obtained with different methods can be directly compared since testing was done under the same conditions.

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 below.

The overall accuracy is

$$\mathrm{Q2} = p/N \tag{1}$$

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

$$C(s) = [p(s)n(s) - u(s)o(s)]/D \tag{2}$$

where D is the normalization factor

$$D = [(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))]^{1/2} \tag{3}$$

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 overpredictions.

The coverage for each discriminated structure s is evaluated as

$$Q(s) = p(s)/[p(s) + u(s)] \tag{4}$$

where $p(s)$ and $u(s)$ are the same as in Equation (2).

The probability of correct predictions $P(s)$ (or accuracy for s) is computed as

$$P(s) = p(s)/[p(s) + o(s)] \tag{5}$$

where $p(s)$ and $o(s)$ are the same as in Equation (2) (ranging from 1 to 0).

The reliability score for each network prediction is also assigned. With one output NN this is obtained by computing

$$\mathrm{Rel}(i) = 20 * \mathrm{abs}(O(i) - 0.5) \tag{6}$$

For computing regression we use the standard correlation (R) and root mean squared standard error (RMSE) values.

3 RESULTS AND DISCUSSION

3.1 Predicting the sign of the protein stability change from sequence

We have previously shown that with an NN-based method over 80% of the mutations in a data set containing 1615 examples were correctly assigned provided that the protein 3D structure was known (Capriotti *et al*., 2004). In this paper we focus on the protein sequence and predict whether a mutation along the sequence increases or decreases the corresponding protein stability without referring to the 3D structure. The results obtained with the different machine learning predictors specifically developed for this task are reported and compared in Table 1. It is interesting to notice that even though the information is only relative to the sequence, an SVM endowed with an RBF kernel reaches an accuracy of 0.77, with a correlation coefficient of 0.42. This finding indicates that a piece of information relevant to the protein folding stability can be traced back to the sequence nearest neighbors of the residue that undergoes mutation. Apparently the RBF kernel is better suited to this task than others. This may indicate that this kernel type properly captures the underlying properties in the residue local environment conducive to the protein stability/instability related also to temperature and pH (routinely physiological) at which mutation occurs.

In Table 2 we show that the best accuracy is reached when the sequence window is 19 residues long. In Table 2 we also test the information pertaining to an infinite window by including the effect

Table 1. Cross-validation performance of the NN and SVM

Method	Q2	$P(+)$	$Q(+)$	$P(-)$	$Q(-)$	C
NeuralNet	0.73	0.39	0.56	0.77	0.87	0.30
SVM-linear	0.67	0.41	0.28	0.73	0.84	0.13
SVM-polynomial	0.73	0.58	0.38	0.77	0.88	0.30
SVM-sigmoid	0.68	0.44	0.27	0.73	0.85	0.15
SVM-RBF	0.77	0.69	0.46	0.79	0.91	0.42

$+$ and $-$: the index is evaluated for positive and negative signs of the protein free energy stability change; for the definition of the different indexes see Section 2.3. The window length for both methods included 19 residues.

Table 2. Cross-validation performance of different window lengths using a RBF kernel

Window	Q2	$P(+)$	$Q(+)$	$P(-)$	$Q(-)$	C
7	0.74	0.58	0.36	0.77	0.89	0.30
11	0.73	0.85	0.12	0.73	0.99	0.25
15	0.76	0.64	0.38	0.78	0.91	0.35
19	0.77	0.69	0.46	0.79	0.91	0.42
23	0.76	0.64	0.44	0.79	0.90	0.38
Whole sequence	0.73	0.59	0.32	0.76	0.90	0.28

For notation see Table 1.

Table 3. Q2 accuracy as a function of the mutated residue type

Native\new	Charged	Polar	Apolar
Charged	0.65 (4%)	0.72 (7%)	0.69 (12%)
Polar	0.57 (5%)	0.76 (5%)	0.77 (13%)
Apolar	0.80 (5%)	0.88 (9%)	0.80 (40%)

Each cell represents a particular type of mutation classified according to chemico-physical properties. Rows account for the wild-type residue (native) and the column positions define the new residues in the mutant proteins (new). In brackets the relative fraction in the protein set (2048) of a given residue type is shown.

of the whole sequence. It is evident that the accuracy diminishes, and this indicates that the whole sequence composition is not as specific as the local sequence environment in terms of determining the sign of the stability change. Also in this case the correlation coefficient is different from random.

The analysis of the SVM accuracy as a function of the chemico-physical properties of the mutations indicates that the protein stability changes involving charged/charged, polar/charged and charged/apolar mutations score lower than those involving apolar/apolar swaps (Table 3), and this suggests that for charged and polar residues at the surface or for charged residues involved in salt-bridges, more information than the local sequence environment is necessary for a high predictive score.

The overall Q2 accuracy is computed as a function of the reliability index (Rel). This identifies a relationship between the reliability value and the predictor accuracy, as shown in Figure 1. The value of the reliability index and its relationship to the prediction accuracy

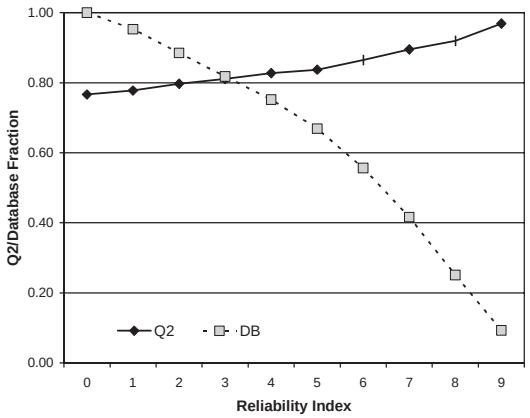

Fig. 1. Q2 accuracy of SVM-RBF as a function of the reliability index (Rel) of the prediction [Equation (6)]. DB is the fraction of the data set with Rel values higher or equal to a given threshold.

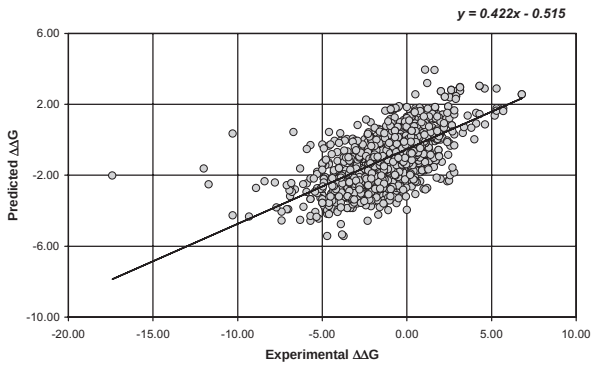

Fig. 2. Regression between predicted and expected values of free energy change upon mutation starting from the protein sequence [$R = 0.62$, RMSE $= 1.45$ $(0.422x - 0.515)$ Kcal/mol].

may help in selecting which mutations are more suited to increasing or decreasing protein stability in a rational computer-aided protein design even at the genomic level.

3.2 Predicting the free energy values of protein stability change from sequence

In specific cases, not only the sign of the mutation but also the exact value of the free energy stability change may be necessary for selecting the mutation type. We have previously shown that coupling machine learning with energy-based methods could provide an excellent solution to this problem (Capriotti *et al*., 2004). However, this is restricted to the small subset of proteins for which a 3D structure is available. Since the aim of this paper is to extend the prediction of stability changes upon mutation to the sequence space, we also implement a SVM that predicts the exact $\Delta\Delta G$ values. This is done using the ν-regression SVM with RBF kernel (libSVM).

In Figure 2 we show the regression between the predicted and the expected $\Delta\Delta G$ values. Predictions are obtained using a 20-fold cross-validation. The R (regression) value is equal to 0.62 with a RMSE of 1.45 Kcal/mol. It should be stressed that this correlation is obtained by starting from the protein sequence and that to our knowledge this is the first method capable of performing the task at

Table 4. Prediction of disease-related mutations

Protein	Mutation	Effect	Predicted stability change	Rel	Experimental $\Delta\Delta G$ (Kcal/mol)	Ref.
Human prion (PRIO_HUMAN)						
	P102L	GSD	Increase	2	0.2 ± 0.6	Apetri *et al*. (2004)
	M129V	Polymorphism	Decrease	6	-0.3 ± 0.5	Liemann and Glockshuber (1999)
	V180I	**GSD**	**Decrease**	**2**	$\mathbf{-0.5 \pm 0.4}$	**Liemann and Glockshuber (1999)**
	T183A	**CJD**	**Decrease**	**6**	$\mathbf{-4.6 \pm 0.7}$	**Liemann and Glockshuber (1999)**
	T190V	Polymorphism	Decrease	2	0.2 ± 0.6	Liemann and Glockshuber (1999)
	F198S	**GSD**	**Decrease**	**7**	$\mathbf{-2.5 \pm 0.4}$	**Liemann and Glockshuber (1999)**
	E200K	CJD	Decrease	5	-0.1 ± 0.6	Liemann and Glockshuber (1999)
	R208H	**CJD**	**Decrease**	**7**	$\mathbf{-1.4 \pm 0.6}$	**Liemann and Glockshuber (1999)**
	V210I	CJD	Decrease	2	-0.3 ± 0.6	Liemann and Glockshuber (1999)
	Q217R	**GSD**	**Increase**	**1**	$\mathbf{-2.1 \pm 0.4}$	**Liemann and Glockshuber (1999)**
	M166V	Polymorphism	Decrease	6	SC(1E1J)	Calzolai *et al*. (2000)
	S170N	Polymorphism	Increase	1	SC(1E1P)	Calzolai *et al*. (2000)
	R220K	Polymorphism	Decrease	7	SC(1FKC)	Calzolai *et al*. (2000)
Transthyretin (TTHY_HUMAN)						
	V50M	**Amyloidosis**	**Decrease**	**6**	$\mathbf{-2.2 \pm 2.4}$	**Shnyrov *et al*. (2000)**
	L75P	**Amyloidosis**	**Decrease**	**5**	$\mathbf{-1.5 \pm 2.3}$	**Shnyrov *et al*. (2000)**
	T139M	Unclassified	Decrease	0	-0.1 ± 2.8	Shnyrov *et al*. (2000)
	T80A	Amyloidosis	Decrease	6	SC(1TSH)	a
	S97Y	Amyloidosis	Increase	2	SC(2TRY)	a
	Y134C	Amyloidosis	Increase	0	SC(1IIK)	a
	V142I	Unclassified	Decrease	2	SC(1TTR)	a

GSD, Gerstmann–Straussler disease; CJD, Creutzfeldt–Jakob disease; Rel, reliability index (see Measure of Accuracy); SC, structural conformational changes determined by comparing the native (1QLX, human prion protein; 1BM7, human transthyretin) with the mutated 3D structures (PDB codes are reported within parentheses); a, derived by comparison between the native structure and the mutated as reported in the PDB files through the SWISSPROT links. Bold lettering indicate the subset of mutations in which the $\Delta\Delta G$ values is $\geq$0.5 Kcal/mol.

hand at this level of efficiency. For this reason we suggest that our approach can be successfully applied when protein structures are not available and thermodynamic data on protein stability need to be analyzed in terms of molecular properties.

3.3 Disease-related single nucleotide polymorphisms and the prediction of protein stability changes

Evidence is accumulating that many disease-causing mutations exert their effects by altering protein folding (Wang and Moult, 2001, 2003; Dobson, 2003; Selkoe, 2003). An interesting application of our method is therefore the prediction of protein stability changes when mutations are known to correlate to diseases.

In Table 4 the predicted thermodynamic data for 20 mutations of the human prion protein and human transthyretin are shown and either compared with the experimental $\Delta\Delta G$ values, when available, or related to conformational changes, when known with atomic resolution. The sign of the stability change is correctly predicted in all cases but two, with a correlation coefficient of 0.42. On this blind test the performance is similar to that on the training/testing set.

It is also interesting to note that the protein stability decrease upon mutation correlates with maladies in 77% of the experimental data. Moreover, if we focus only on the subset in which the $\Delta\Delta G$ changes are $\geq$0.5 Kcal/mol, all the mutations correspond to diseases. On this subset of experimental data, our predictor fails only in one case to assign the correct $\Delta\Delta G$ sign. However, if we sort the predictions by the reliability index value, all the predictions made with reliability index >2 agree with the experimental data. The results of this test are therefore in agreement with the general idea that defective protein folding is one of the causes of human diseases and suggest also a possible application of this predictor to correlate single nucleotide polymorphisms and diseases related to protein instability.

ACKNOWLEDGEMENTS

This work was supported by the following grants: “Hydrolases from Thermophiles: Structure, Function and Homologous and Heterologous Expression” of the Ministero della Istruzione dell’Università e della Ricerca (MIUR); a PNR 2001–2003 (FIRB art 8) project on postgenomics to R.C. E.C. is supported by a grant from the European Union’s VI Framework Programme for the BioSapiens Network of Excellence project. P.F. acknowledges an MIUR grant on proteases.

Conflict of Interest: none declared.

REFERENCES

Apetri,A.C. *et al*. (2004) The effect of disease-associated mutations on the folding pathway of human prion protein. *J. Biol. Chem.*, **279**, 31048–31052.

Bava,K.A. *et al*. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **32**, D120–D121.

Calzolai,L. *et al*. (2000) NMR structures of three single-residue variants of the human prion protein. *Proc. Natl Acad. Sci.* USA, **97**, 8340–8345.

Capriotti,E. *et al*. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **20** (suppl. 1), I63–I68.

Daggett,V. and Fersht A.R. (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci*., **28**, 18–25.

Dobson,C.M. (2003) Protein folding and misfolding. *Nature*, **426**, 884–890.

Funahashi,J. *et al*. (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 Rooman,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.

Guerois,R. *et al*. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.

Kwasigroch,J.M. *et al*. (2002) PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics*, **18**, 1701–1702.

Liemann,S. and Glockshuber,R. (1999) Influence of amino acid substitutions related to inherited human prion diseases on the thermodynamic stability of the cellular prion protein. *Biochemistry*, **38**, 3258–3267.

Pitera,J.W. and Kollman,P.A. (2000) Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. *Proteins*, **41**, 385–397.

Prevost,M. *et al*. (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.

Shnyrov,V.L. *et al*. (2000) Comparative calorimetric study of non-amyloidogenic and amyloidogenic variants of the homotetrameric protein transthyretin. *Biophys. Chem.*, **88**, 61–67.

Selkoe,D.J. (2003) Folding proteins in fatal ways. *Nature*, **426**, 900–904.

Topham,C.M. *et al*. (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **10**, 7–21.

Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263–270.

Wang,Z. and Moult,J. (2003) Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins*, **53**, 748–757.

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.