Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604
Vol 2, Issue 2, June-2011, pp 155-160 http://www.bipublication.com

PLS FOR PREDICTION OF THERMAL STABILITY OF PROTEIN MUTANTS

Sekhar Talluri
Department of Biotechnology, GIT, GITAM University

ABSTRACT:

Prediction of the stability of proteins is a critical step in the goal of protein design. The Partial Least Squares (PLS) regression method was used to predict the thermal stability of protein variants produced by single point mutations. The efficacy of this method was tested by using data on changes in free energy of unfolding (ddG) of mutations of phage T4 lysozyme. These data were obtained from the ProTherm database. PLS regression models were built by using physico-chemical descriptors. Structural information can also be incorporated into these models, if it is available. The models obtained in this manner were able to fit the experimental data on free energy of unfolding with an rms error of 0.42 kcal/mol and a two state classification accuracy of 89.9%. The magnitude of the rms error for these models is comparable to the magnitude of the expected errors in the experimental data. However, it was observed that less complex models, obtained from PLS models of lower rank, are more robust. The predictive rms error for these robust models were 1.2 kcal/mol (without structural information) and 1.1 kcal/mol (with structural information).

Keywords: *Thermal stability; Partial least squares; point mutation*

[I] INTRODUCTION

Prediction of protein stability is a critical requirement for protein design. There are three classes of methods for prediction of thermal stability of mutant proteins [1]: physical, knowledge based and empirical. Physical methods for prediction of thermal stability are based on statistical mechanics and may use molecular mechanics and/or quantum mechanics for computation of the free energy of unfolding. Methods based on free energy perturbation theory are generally regarded as the most accurate. However, these methods are computationally intensive and require considerable user expertise. A variety of knowledge based methods are also available [2-14]. The primary advantages of the knowledge based methods are speed and ease of use. Empirical methods combine information from molecular mechanics potentials and statistical analysis [15]. The computational effort involved in empirical methods is generally higher than that of knowledge based methods. A recent evaluation of the knowledge-based and empirical methods indicates that the prediction accuracies of these methods [16] were approximately 60%.

The basic assumption of the method described here is that the physico-chemical properties of the constituent amino acids and their specific interactions determine the thermal stability of a protein. A large number of physico-chemical descriptors of amino acids have been described and their role in prediction of thermal stability has been assessed both experimentally [17] and computationally [18]. Therefore, the thermal stability of proteins can be predicted by a proper choice of the physico-chemical descriptors and by determination of their relative contribution [19]. A very large number of choices are possible, since a large number of physico-chemical descriptors are available and the contribution of a physico-chemical descriptor, such as polarity, varies in a sequence and

structure dependent manner. In addition, if the three-dimensional structure is available, a number of scoring functions and molecular mechanics based potential energy functions are available for estimating the energetic contribution of interactions between the constituent amino acids of a protein.

Subsets of the available descriptors may be used to construct predictive models by using multiple linear regression. However, the number of such models is prohibitively high due to the large number of descriptors available (>100). Models based on principal components analysis or partial least squares analyze the variance/covariance to identify descriptors that have the best predictive ability. In principal component analysis the variance associated with descriptors is used to guide the model building. In the partial least squares method, the covariance of the descriptors and the dependent variables to be predicted is considered – this is expected to produce models with better predictive ability [20]. The partial least squares method has been used extensively for the establishment of structure-property relationships [20] for lead optimization during drug development, and it is one of the most effective methods for chemometrics, but it has not been used earlier for prediction of thermal stability of proteins. The objective of this study is to evaluate the effectiveness of the partial least square method for building predictive models of thermal stability.

[II] MATERIALS AND METHODS

2.1. Data sets

The data on thermal stability of T4 lysozyme was obtained from the Protherm database [21]. The data set referred here as prothermlysot4 (440 entries) was obtained by removing duplicate entries. An entry was considered as duplicate if another entry having the same residue number and experimental conditions was already included in the dataset. The prothermlysot4 dataset was pruned further to ensure that only one entry was included per mutation by removing entries corresponding to different experiments on the same mutant - this dataset will be referred to as nrdlysot4 (244 entries). Data sets for other proteins were constructed in a similar manner.

2.2. Descriptors

Descriptors were either obtained from tables of amino acid properties or they were computed. Amino acid properties to be used as descriptors [22] for model construction were obtained from Genomenet (http://www. genome.ad.jp/aaindex/). The following properties were found to be particularly useful: Unfolding Gibbs energy in water, Transfer free energy from octanol to water, Graph shape index.

2.2.1. pH dependent charge and electrostatic interaction energy

The experimental data on thermal stability, obtained from ProTherm, includes entries with several different values of pH. The expected charge on amino acids with titratable side chains, at the pH of the experiment, was calculated by using the pKa values and the Henderson-Hasselbach equation. These computed charges were then used to estimate the electrostatic interaction energy for models based on structural data. In addition, the experimentally determined HPLC retention coefficients tabulated for each of the amino acids at pH 2.1 and pH 7.4 were used to predict the retention coefficients at the pH used for determination of the free energy of unfolding of the mutant. The pH dependent retention coefficients were estimated from the calculated fraction of the acid and base form of the amino acids at the specified pH value. These pH dependent retention coefficients, representing hydrophobicity, were one of the physico-chemical parameters used as descriptors in some models.

2.2.2. Modeller

The Modeller program package [23] was used to build models of the mutant proteins using the native protein structure as a template. For each

model, the pairwise energy, the total energy, the molecular probability density function and the DOPE score [23] were calculated. These values were used as descriptors in some of the PLS models. These values were calculated once, and stored, to ensure rapid calculation of thermal stability.

2.3. Model construction

The partial least squares regression coefficients were determined by using the nonlinear iterative Partial Least Squares algorithm [20] (NIPALS). The subroutine for implementing this algorithm was obtained from a QSAR program package [24]. The implementation was designed to allow a variety of physico-chemical and interaction functions to be used as descriptors.

2.4. Model evaluation

The models were optimized and evaluated based on their ability to reproduce the quantitative changes in the free energy of unfolding of the mutant proteins. The predictive ability was assessed by using the calculated rms error between the experimental value and the predicted value of the change in the free energy of unfolding. However, other measures of prediction, such as two-state accuracy, three-state accuracy and Matthews correlation coefficient were also calculated.

The two-state accuracy refers to the prediction of the sign of the difference in the free energy of unfolding of the mutant and the native protein, i.e., it predicts whether a mutation is stabilizing or destabilizing compared to the native protein.

Three-state accuracy measures the ability to classify the mutations into three classes: stabilizing, destabilizing and neutral. A mutation is deemed to be neutral if the free energy of unfolding is within +/-0.5 kcal/mol of the free energy of unfolding of the native protein.

The predictive accuracy of the results was assessed by using Leave-one-out (LOO) cross validation method. In addition, 10-fold cross validation was implemented by constructing 10 data sets from the original data; in each of these, 10% of the entries were set aside (randomly) as test data and the PLS regression coefficients were calculated from the remaining 90%. The predictive accuracy was assessed by averaging the prediction results for the 10 test data sets.

[III] RESULTS

Partial least squares regression maximizes the covariance between a set of descriptors and the dependent variables (experimentally determined values of thermal stability). The descriptors were chosen carefully based on their physical significance. However, the PLS regression method is tolerant of descriptors that contain redundant information, because the PLS regression model determines the optimal weights for the descriptors.

The results obtained for three PLS regression models for prediction of thermal stability of the nrdlysot4 dataset are summarized in Table 1.

3.1. Model 1

Model 1 was constructed by using a set of four physico-chemical parameters and it does not make use of any structural information. The four physico-chemical properties of the aminoacids used as descriptors in this model were: Graph shape index, Transfer free energy from octanol to water, Unfolding Gibbs energy in water at pH 7.0 and computed charge.

[Table-1] Models for prediction of thermal stability of mutant proteins.

	Model1 (Best model without structural data)	Model2 (Prediction model without structural data)	Model3 (Prediction model with structural data)
RMS error	**0.42 kcal/mol**	0.93 kcal/mol	0.86 kcal/mol
RMS error (CV using LOO)	-	1.27 kcal/mol	1.11 kcal/mol
Predictive RMS error (10-fold CV)	-	**1.22 kcal/mol**	**1.12 kcal/mol**
Accuracy (2-State)	**89.93%**	74.70%	81.12%
Predictive Accuracy (2-state) (CV using LOO)	74.49%	74.60%	76.18%
Accuracy (3-State)	83.75%	61.49%	67.26%
Predictive Accuracy (3-state) (CV using LOO)	60.23%	55.91%	62.28%
R^2	**0.92**	0.61	0.67
Predictive R^2 (10-fold CV)	-	0.28	0.41
Q^2	-	0.27	0.44
Matthews correlation coefficient	0.74	0.28	0.5
Matthews correlation coefficient (CV using LOO)	0.35	0.28	0.36

The two-state prediction accuracy of Model 1 is 89.9%. Model 1 was able to predict the free energy of unfolding with an RMS error of 0.42 kcal/mol. This prediction error is comparable to the estimated error for experimental determination of free energy of unfolding (> 0.37 kcal/mol). For example, the experimental values of free energy difference for unfolding for the A82P mutation of T4 lysozyme at pH2 were observed to be -0.07 kcal/mol using DSC and +0.3 kcal/mol by using CD (Protherm entry numbers 13728 and 13616). However, Model 1 is not very robust, and it was observed that there was a considerable decrease in accuracy of prediction upon cross-validation.

3.2. Model 2

Model 2 was also constructed from the same set of physico-chemical parameters as Model 1, but Model 2 was obtained from a set of PLS coefficients corresponding to a lower rank. The rms error for this model is considerably higher (0.93 kcal/mol) than that of Model 1. However, this model is more robust and it was deemed suitable for prediction of thermal stability.

3.3. Model 3

Model 3 was constructed by using a set of three physico-chemical parameters (free energy of unfolding, transfer free energy from octanol to water and computed charge) and information regarding the pairwise interaction energy, combined energy, molecular probability density function and the DOPE score from structural models constructed by using Modeller. The rms error for this model is 0.86 kcal/mol and the predictive rms with 10-fold cross-validation is 1.1 kcal/mol. The two-state predictive accuracy of this model is 76% using LOO cross-validation.

[IV] DISCUSSION

Most of the semi-empirical and knowledge based methods for prediction of thermal stability of protein do not require information regarding thermal stability of single-point mutations of the protein of interest in the training data set. However, specific models have been constructed, such as a model to predict the stability of human lysozyme mutants based on the use of heat capacity curves [25]. Such models are expected to be useful for planning experimental work on new mutations [25].

The current implementation of the PLS model is based on the use of training data sets consisting of single-point mutations of the protein of interest. The more widespread adoption of large scale methods for production of mutants such as alanine-scanning and automation of methods for determination of mid-point of thermal transition

indicate that thermal stability data would be more readily available in the near future.

The rms error of fit for the PLS model without structural information (Model 1), is better than that of any other prediction method that does not require structural information. However, the accuracy of Model 3, measured in terms of rms error, appears to be comparable to the best methods for prediction of thermal stability that utililize structural information. The two-state accuracy of Model 3 is slightly lower than that reported for the best methods that use structural information [13,14]. However, the PLS models described here were not optimized as classifiers and in addition the outliers were not eliminated. The prediction accuracy can be be improved by removing a few outliers from the training/test data sets – outliers were not eliminated in the results described in Table I. Direct comparison of these results with the results published with other computational methods is not possible because the published methods differ from each other considerably in the selection of the test data sets, treatment of outliers and the methods used for cross validation.

The current implementation of the PLS regression model is applicable for variants consisting of a change in a single amino acid; however, it can be easily extended to proteins having two or more mutations. If thermal stability data is available for one protein and the structural data is available for a homologous protein, then comparative modeling can be useful for prediction of thermal stability [26]. The current implementation of the PLS model can be modified to accommodate such extensions, because the energetic contributions required for the current implementation are estimated by using the Modeller program package which is designed for comparative modeling. Therefore, the current implementation of the PLS regression method for prediction of thermal stability of single-point mutations can be regarded as a first step in the use of physico-chemical parameters (and molecular mechanics derived interaction energies) for prediction of the thermal stability of homologous proteins.

[V] CONCLUSION

The ability of the Partial Least Squares regression method to predict the thermal stability of variants of proteins produced by single point mutations has been assessed. This method was tested by using data on changes in free energy of unfolding (ddG) of mutations of phage T4 lysozyme. The model obtained in this manner was able to fit the experimental data on free energy of unfolding with an rms error of 0.42 kcal/mol and a two state classification accuracy of 89.9%. The magnitude of the rms error for this model is comparable to the magnitude of the expected errors in the experimental data. PLS models of lower rank were more robust, however slightly higher predictive rms errors of 1.2 kcal/mol (without structural information) and 1.1 kcal/mol (with structural information), were observed. There is only a small difference between the predictive rms error for models with and without structural information. This is because of the ability of the PLS method to effectively estimate the contribution of a physico-chemical parameter to each site in the protein using the information regarding the thermal stability of mutants in the training data set. However, if the size of the training data set is much smaller than the one used in this study, the relative effectiveness of the structure independent model is expected to decrease. The addition of structural data provides the additional information necessary for maintaining the accuracy of the model. Therefore, it is recommended that the model which incorporates structural information (Model 3) should be used if the three-dimensional structure of the native protein is available.

The rms errors of the models reported here are comparable to the rms errors of the most effective computational methods that have been used for prediction of thermal stability of protein mutants [14,15]. Furthermore, the method reported here

has the potential for extension to related problems such as prediction of the thermal stability of homologous proteins.

ACKNOWLEDGEMENT

The subroutine for implementation of the PLS regression algorithm was obtained from the QSAR program, which was made available as an open source program by J. Ponder, University of Washington at St. Louis.

REFERENCES

[1] Lazarides T and Karplus M. [2000] Effective energy functions for protein structure prediction, *Current opinion in structural biology* **10**:139-145.

[2] Capriotti E, Fariselli P and Casadio R. [2004] A neural-network-based method for predicting protein stability, *Bioinformatics* **20**:i63-i68.

[3] Emidio Capriotti, Piero Fariselli and Rita Casadio, [2005] I-Mutant2.0: predicting stability changes upon mutation, *Nucleic Acids Research* **33**:W306–W310.

[4] Christian H and Schomburg D. [2005] Prediction of protein thermostability with a direction- and distance-depend, *Protein Science* **14**:2682–2692.

[5] Parthiban V, Gromiha MM and Schomburg D, [2006] CUPSAT: prediction of protein stability upon point mutations, *Nucleic Acids Research* **34**:W239–W242.

[6] Demenkov P, Aman E. and Ivanisenko V. [2006] Prediction of the changes in thermodynamic stability of proteins, *Biophysics* **51**:49-53.

[7] Saraboji K, Gromiha MM and Ponnuswamy MN. [2006] Average assignment method for predicting the stability of protein mutants, *Biopolymers* **82**:80-92.

[8] Masso M and Vaissman II. [2008] Accurate prediction of stability changes in protein mutants, *Bioinformatics* **24**:2002-2009.

[9] Huang L-T, Gromiha MM and Ho SY. [2007] iPTREE-STAB: interpretable decision tree based method, *Bioinformatics* **23**:1292-1293.

[10] Gromiha MM. [2007] Prediction of protein stability upon point mutations, *Biochemical Society Transactions* **35**:1569-1573.

[11] Cheng J, Randall A and Baldi P. [2006] Prediction of protein stability changes for single-site mutations, *Proteins* **62**:1125-1132.

[12] Huang LT, Gromiha MM and Ho SY. [2007] Sequence analysis and rule development of predicting protein stability, *Journal of molecular modeling* **13**:879-890.

[13] Huang L-T and Gromiha MM. [2009] Reliable prediction of protein thermostability change upon double mutation, *Bioinformatics* **25**:2181-2187.

[14] Teng S, Srivastava AK and Wang J. [2010] Sequence feature-based prediction of protein stability changes, *BMC Genomics* **11**:S5.

[15] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F and Serrano L.[2004] The FoldX web server: an online force field, *Nucleic Acids Research* **33**:W382-388.

[16] Vihinen M and Khan S. [2010] Performance of protein stability predictors, *Human mutation* **31**:675-685.

[17] Mooers BHM, Baase WA, Wray JW and Matthews BW. [2009] Contributions of all 20 amino acids at site 96 to the stability of lysozyme, *Protein Sscience* **18**:871-880.

[18] Kumar S, Tsai CJ and Nussinov R. [2000] Factors enhancing protein thermostability, *Protein Engineering* **13**:179-191.

[19] Kang S, Chen G and Xiao G. [2009] Robust prediction of mutation-induced protein stability change, *Protein engineering design and selection* **22**:75-83.

[20] Rannar S, Lindgren F, Geladi P and Wold S, A. [1994] PLS Kernel Algorithm for Data Sets with many Variables and fewer Objects, *Journal of Chemometrics* **8**:111-125.

[21] Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H and Sarai A. [2006] *Nucleic acids research* **34**:D204-206.

[22] Kawashima S, Ogata H and Kanehisa M. [1999] AAindex: amino acid index database, *Nucleic acids research* **27**:368-369.

[23] Sali A and Blundell TL. [1993] Comparative protein modelling by satisfaction of spatial restraints, *Journal of Molecular Biology* **234**:779-815.

[24] Fedders M and Ponder J. [1996] QSAR program package, University of Washington, St. Louis.

[25] Verma D, Jacobs DJ, Livesay DR. [2010] Predicting the Melting Point of Human C-Type Lysozyme Mutants, *Current protein & peptide science* **11**:562-572.

[26] Bloom JD and Glassman MJ. [2009] Inferring stabilizing mutations from protein phylogenies, *PLOS computational biology* **5**:e1000349.