EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models

Lukas Folkman and Yaoqi Zhou

Abstract

Our submission to TPMT/PTEN stability prediction challenge in CAGI 5 was based on our method, named EASE-MM, which was published previously [1]. The submitted predictions were realized using our publicly available web-server: `http://sparks-lab.org/server/ease`. This document describes the method in detail. The text used in this document was extracted in full from our publication.

1 Materials and Methods

We have built EASE-MM (Figure 1), which comprises five specialised models to predict $\Delta\Delta G_u$ of mutations in residues located in different secondary structure (SS) elements (helix, sheet, or coil) and with different levels of accessible surface area (ASA) (exposed or buried with a 25% threshold). The final prediction is the average of $\Delta\Delta G_u$ predicted with two models, one selected based on the predicted SS and the other based on the predicted ASA of the mutation site. We used a dataset of 1676 mutations (S1676) to design our method and estimate its performance using 10-fold cross-validation. Next, we employed two independent datasets of 543 and 236 mutations (S543 and S236, respectively) to confirm the robust performance of our method. Both datasets had a sequence identity $< 25\%$ to the dataset used for the design and training of our method. Finally, we studied the relationship between disease-causing germline SNVs and $\Delta\Delta G_u$ predicted with EASE-MM.

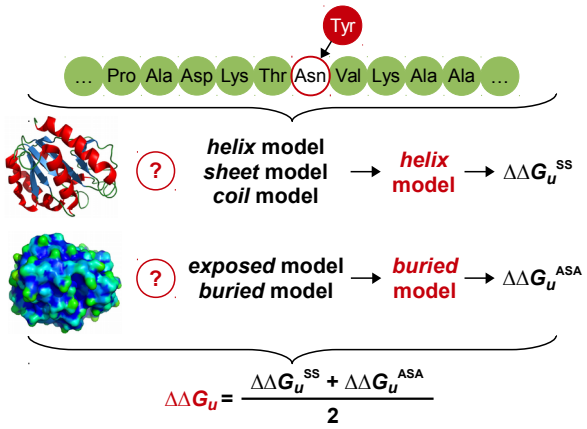

Figure 1: EASE-MM calculates the stability change ($\Delta\Delta G_u$) as the average of $\Delta\Delta G_u$ predicted with two distinct models chosen based on the *predicted* secondary structure and accessible surface area of the mutation site.

1.1 Datasets

We used several different datasets to design, validate, and independently test our method. We used the S1676 (1676 mutations in 70 proteins) and S236 (236 mutations in 23 proteins) datasets compiled in Folkman *et al.* [2] from ProTherm [3] (version February 2013). ProTherm defines a stability change as the difference in the unfolding free energy: $\Delta\Delta G_u[\mathrm{kcal\,mol^{-1}}] = \Delta G_u(mutant) - \Delta G_u(wild\text{-}type)$. Thus, destabilising mutations yield $\Delta\Delta G_u < 0$. We verified all records in ProTherm and corrected incorrect entries according to the original publications. Next, we removed all duplicate entries of the same amino acid substitutions (*e.g.*, different concentrations of chemicals). If several measurements of the same mutation under the same experimental conditions were present, we averaged $\Delta\Delta G_u$. If several measurements of the same mutation under different experimental conditions were present, we kept only the measurement closest to the physiological pH 7. The S1676 dataset was used to design our method and optimise all parameters. S1676 and S236 are mutually independent and do not share proteins with $\geq 25\%$ sequence identity. Therefore, we used S236 for independent testing and comparison with related work. Another independent test set, S543 (543 mutations in 55 proteins), was compiled as a subset of the 2648 mutations from Dehouck *et al.* [4]. S543 has $< 25\%$ sequence identity to both S1676 and S236.

Finally, to study the relationship between the predicted $\Delta\Delta G_u$ and human germline non-synonymous SNVs, we compiled a dataset of 10,511 disease-causing (2201 proteins) and 278,760 putatively neutral (20,096 proteins) SNVs from ClinVar [5] and 1000 Genomes Project [6], respectively. For the distribution analysis, we considered the subset comprising 50,910 putatively neutral SNVs (14,113 proteins) with AF $\geq 1\%$.

1.2 Predictive features

We employed three types of predictive features in our method: evolutionary conservation, amino acid parameters, and predicted structural properties. To estimate evolutionary conservation of the mutation site, we used three iterations of PSI-BLAST [7] on the NCBI non-redundant database with an *e*-value threshold of 10^{-3}. From the PSSM generated with PSI-BLAST, we extracted the probability of the wild-type (PSSM_{wt}) and mutant (PSSM_{mt}) amino acids at the mutation site. We implemented two different features, PSSM_{wt} and $\Delta\,\mathrm{PSSM} = \mathrm{PSSM}_{mt} - \mathrm{PSSM}_{wt}$. We also included a feature encoding the overall conservation of the mutation site as property entropy (PE) with respect to six sets grouping amino acids based on their chemical properties as aliphatic (A, V, L, I, M, C), aromatic (F, W, Y, H), polar (S, T, N, Q), positive (K, R), negative (D, E), and special (G, P) [8]. The property entropy was calculated from a multiple sequence alignment of the 30 most similar sequences from the NCBI non-redundant database ranked by *e*-value with PSI-BLAST (using 100, 500, or all sequences resulted in a lower correlation with $\Delta\Delta G_u$, S1676 dataset). We used the implementation from Capra and Singh [9] to calculate the property entropy based on the following equation:

$$\mathrm{PE}(msa_i) = 1 - \left(- \frac{\sum_{g \in G} p(msa_i, g) \times \log p(msa_i, g)}{\log |msa_i|} \right),$$

$$p(msa_i, g) = \sum_{aa \in g} p(msa_i, aa),$$

where msa_i is the i-th column of the multiple sequence alignment msa, G is the set of the defined property groups, and $p(msa_i, g)$ is the probability of the property group g at msa_i, which is equal to the sum of probabilities of the amino acid types (aa) belonging to g.

Different amino acid parameters have been used for the prediction of stability changes [10, 11, 12, 13]. We adopted a total of 11 amino acid parameters: hydrophobicity, volume, polarisability, isoelectric point, helix tendency, sheet tendency, and a steric parameter (graph shape index) from Meiler *et al.* [14]; compressibility, bulkiness, and equilibrium constant with reference to the ionisation property of COOH

group from Gromiha *et al.* [15]; and flexibility from Vihinen *et al.* [16]. For each amino acid parameter (AAP), we calculated $\Delta\,\mathrm{AAP} = \mathrm{AAP}_{mt} - \mathrm{AAP}_{wt}$, where AAP_{mt} and AAP_{wt} denote the value of the given AAP for the mutant and wild-type amino acids, respectively.

Finally, we considered five structural features predicted from the protein sequence: rASA, helix, sheet, coil, and disorder probabilities. The rASA and SS probabilities were predicted using SPIDER [17]. The disorder probability was calculated using SPINE-D [18].

1.3 Feature selection and multiple models

To build the five models employed by EASE-MM, we partitioned the S1676 training dataset according to SS (helix, sheet, and coil) and ASA (buried or exposed with a 25% threshold). SS and ASA were predicted from the protein sequence with SPIDER [17].

A unique set of features was identified for each of the five SVM models using the SFFS algorithm [19]. SFFS starts with an empty set of features S_0 and iteratively searches for a better set of features in two steps. First, the best feature f is selected as the one for which $S_i = S_{i-1} \cup \{f\}$ yields the lowest RMSE. Second, features f^* for which $S_i - \{f^*\}$ yields a lower RMSE than S_{i-1} are iteratively removed. Thus, the number of features in S is not monotonously increasing because the search is 'floating' up and down.

1.4 Training and evaluation

We employed the S1676 dataset to design our method, perform feature selection, and optimise all parameters using the *unseen-protein* 10-fold cross-validation. The *unseen-protein* cross-validation is used to avoid over-fitting on specific proteins by splitting the dataset into cross-validation folds so that all mutations of a cluster of similar proteins ($\geq 25\%$ sequence identity) are always contained within a single fold. These clusters were identified with Blastclust [20]. To devise a robust estimate of the prediction performance, we replicated the cross-validation procedure 100 times with randomly re-generated folds and averaged the results.

We implemented EASE-MM with ϵ-SVR (support vector regression) and radial basis function (RBF) kernel using the LibSVM [21] library. For ϵ-SVR, we optimised three parameters (C, γ, and ϵ) using a grid search in the range of $C \in \{2^{-1}, 2^0, \ldots, 2^6\}$, $\gamma \in \{2^{-8}, 2^{-7}, \ldots, 2^0\}$, and $\epsilon \in \{2^{-8}, 2^{-7}, \ldots, 2^{-1}\}$. Then, each model was trained on S1676 and tested independently using S543 and S236 to confirm that our approach did not result in over-fitting. Importantly, S543 and S236 did not share similar sequences ($\geq 25\%$ sequence identity) with S1676 and were not used during the design, feature selection, or parameter optimisation in any way. Furthermore, S543 and S236 were disjoint with $< 25\%$ sequence identity. The performance of EASE-MM was assessed in terms of Pearson correlation coefficient (r) and root mean square error (RMSE):

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}},$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (x_i - y_i)^2}.$$

Finally, we re-optimised the ϵ-SVR parameters and trained the final EASE-MM models, which are deployed on our web-server, using a joint S1676+S236 dataset in order to maximise the size of the training data.

References

[1] L. Folkman, B. Stantic, A. Sattar, Y. Zhou, Ease-mm: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models, Journal of Molecular Biology 428 (6) (2016) 1394–1405.

[2] L. Folkman, B. Stantic, A. Sattar, Feature-based multiple models improve classification of mutation-induced stability changes, BMC Genomics 15 (Suppl 4) (2014) S6.

[3] M. Kumar, K. Bava, M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, A. Sarai, ProTherm and ProNIT: Thermodynamic databases for proteins and protein–nucleic acid interactions, Nucleic Acids Research 34 (Suppl 1) (2006) D204.

[4] Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, M. Rooman, Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0, Bioinformatics 25 (19) (2009) 2537.

[5] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, D. R. Maglott, Clinvar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Research 42 (D1) (2014) D980–D985.

[6] The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, Nature 467 (7319) (2010) 1061–1073.

[7] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research 25 (17) (1997) 3389.

[8] L. A. Mirny, E. I. Shakhnovich, Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function, Journal of Molecular Biology 291 (1) (1999) 177–196.

[9] J. A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics 23 (15) (2007) 1875–1882.

[10] L. Huang, K. Saraboji, S. Ho, S. Hwang, M. Ponnuswamy, M. Gromiha, Prediction of protein mutant stability using classification and regression tool, Biophysical Chemistry 125 (2–3) (2007) 462–470.

[11] S. Kang, G. Chen, G. Xiao, Robust prediction of mutation-induced protein stability change by property encoding of amino acids, Protein Engineering Design and Selection 22 (2) (2009) 75.

[12] B. Shen, J. Bai, M. Vihinen, Physicochemical feature-based classification of amino acid mutations, Protein Engineering Design and Selection 21 (1) (2008) 37–44.

[13] S. Teng, A. Srivastava, L. Wang, Sequence feature-based prediction of protein stability changes upon amino acid substitutions, BMC Genomics 11 (Suppl 2) (2010) S5.

[14] J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, Molecular modeling annual 7 (9) (2001) 360–369.

[15] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Relationship between amino acid properties and protein stability: buried mutations, Journal of Protein Chemistry 18 (5) (1999) 565–578.

[16] M. Vihinen, E. Torkkila, P. Riikonen, Accuracy of protein flexibility predictions, Proteins: Structure, Function, and Bioinformatics 19 (2) (1994) 141–149.

[17] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Scientific Reports 5 (2015) 11476.

[18] T. Zhang, E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky, Y. Zhou, SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method, Journal of Biomolecular Structure and Dynamics 29 (4) (2012) 799–813.

[19] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (11) (1994) 1119–1125.

[20] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215 (3) (1990) 403–410.

[21] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 27:1–27:27.