Drug Designing: Open Access
ISSN: 2169-0138

Fang, Drug Des 2015, 4:3
DOI: 10.4172/2169-0138.1000e130

Editorial Open Access

Reliability of Machine Learning Based Algorithms for Designing Protein Drugs with Enhanced Stability

Jianwen Fang*

Biometrics Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 9609 Medical Center Dr., Rockville, MD 20850, USA

Native proteins are often susceptible to physical and chemical degradation because many of them are only marginally stable under both normal physiological and storage conditions. Therefore designing protein drugs with enhanced stability through predicted stabilizing mutations using computational methods has attracted increasing interest in recent years [1]. Data mining technologies employing various machine learning (ML) algorithms have been explored for such a purpose [2-6]. In general ML approaches involve training predictive models based on available experimental data using features (properties) supposedly relevant to protein stability. ML algorithms such as support vector machines [7], neuronal networks [8], and multiple regression and classification techniques [9,10], have been employed in this research area. Substitution types, secondary structures, solvent accessibility, and the amino acid composition of neighboring residues have been commonly used as features in models for predicting stabilizing mutations. The ML approaches hold great promises because they may reveal subtle patterns governing mutation induced stability changes and protein stability in general. Therefore, they not only have significant practical value but also are of great theoretical interest. We and others discovered, however, that some of published methods suffer from the over-fitting problem and suggested the problem can be easily detected by using hypothetical reverse mutations [3,4,11]. Nevertheless, recently we have found, disappointedly, several recent publications still suffer from the same problem [2,7]. In this editorial, we want to alert the research community the pitfall and offer thoughts for moving the field forward.

Protein stability changes upon mutations are often measured experimentally through changes in the folding free energies ($\Delta\Delta G$) or melting temperature (ΔT_m) between wild type proteins and their mutants. Because free energy and temperature are thermodynamic parameters and thus state functions [12], the $\Delta\Delta G$ and ΔT_m of a mutation from a wild type protein to its mutant (WT➔MT) *always* equal the negated $\Delta\Delta G$ (or ΔTm) of the reverse mutation (MT ➔ WT), i.e.,

$$\Delta\Delta G_{WT \rightarrow MT} \equiv - \Delta\Delta G_{MT \rightarrow WT} \quad (1)$$

$$\Delta T_{m\ WT \rightarrow MT} \equiv - \Delta T_{m\ MT \rightarrow WT} \quad (2)$$

Therefore, hypothetical reversed mutations provide a convenient method to test whether a predictor is robust. To perform the test, we identified 48 mutations in ProTherm database [13] for which both wild type and mutant protein structures were available. Therefore, both forward and reverse mutations can be tested. Unfortunately, as we show in the (Table 1), the performance of the all four tested algorithms, including recent mCSM and DUEL (published in 2014), to predict the reverse mutations is far worse than the forward mutations and, in fact, close to random assignment. Therefore all these tested methods suffer from the over-fitting problem, as the forward mutations were likely used in the training (since all methods used the ProTherm data) but the hypothetical reversed mutations were not.

We suggest that the main causes for the over fitting problem include that the numbers of training cases were too small and also the features used in the models were not sufficiently informative for the task. Almost all models were built on the mutation data collected in ProTherm, a public database devoted to document thermodynamic parameters for wild type and mutant proteins [13]. Often, only a few thousands of data points were used in the training and test of predictive models. These numbers are rather small if one considers the fact that there are 380 different types of single mutations. The situation is further exacerbated by the fact that experiments could be performed at different conditions (e.g., pH and temperature) that may significantly affect protein stability [7].

Mutation directions		WT➔MT	MT➔WT
mCSM [7]	AUC	0.978	0.498
	Accuracy	0.819	0.307
	R	0.667	0.038
DUEL [2]	AUC	0.878	0.475
	Accuracy	0.843	0.370
	R	0.667	0.057
muPro [5]	AUC	0.978	0.441
	Accuracy	0.969	0.284
	R	0.972	0.001
iMutant [6]	AUC	0.975	0.540
	Accuracy	0.898	0.307
	R	0.941	0.040

Table 1: The performance of $\Delta\Delta G$ prediction by various algorithms for mutations and hypothetical reversed mutations. AUC: area under ROC curve, 1 is perfect and 0.5 is random; R: correlation coefficient, 1/-1 is perfect and 0 is random; accuracy: the accuracy of classifying mutations into stabilizing mutations ($\Delta\Delta G$>0) and destabilizing mutations ($\Delta\Delta G$<0), 1 is perfect and 0.5 is random.

Another critical requirement for all ML methods to work properly is a collection of informative features. In our opinion it is very challenging to generate informative features for predicting protein stability changes upon mutations. The energy needed to stabilize/destabilize a protein is quite small. Most folded globular proteins are only stable by 20-60 KJ/mol, relative to their unfolded forms. Mutation induced stability changes are usually at an even smaller scale. For example, the stabilizing mutations and destabilizing mutations archived in ProThermo cause -6.75 KJ/mol and 4.65 KJ/mol differences, respectively, in average between wild type proteins and their corresponding mutants.

***Corresponding author:** Jianwen Fang, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 9609 Medical Center Dr., Rockville, MD 20850, USA, Tel: 240-276-7672; E-mail: jianwen.fang@nih.gov

Received December 19, 2015; **Accepted** December 21, 2015; **Published** December 28, 2015

Citation: Fang J (2015) Reliability of Machine Learning Based Algorithms for Designing Protein Drugs with Enhanced Stability. Drug Des 4: e130. doi:10.4172/2169-0138.1000e130

Considering the energy of one hydrogen bond is in the range of 5 to 50 KJ/mol, a net gain or lose of a hydrogen bond of a mutant over its wild type counterpart can significantly (de)stabilizes the mutant. There are usually hundreds of hydrogen bonds formed in a typical protein because it was estimated that the number of hydrogen bonds in a folded protein can be at least two per amino acid residues [14]. Besides, the strength of hydrogen bond highly depends on the distances and angles between the three involved atoms. Therefore, the margin of error of predicting mutation induced stability is so small that unlikely it can be accurately predicted based on mainly the types and counts of the amino acids residues around the mutation sites, spatially or sequentially, because these features are not sensitive to the local changes induced by mutations. For example, the values of these features for a forward mutation and its corresponding reverse mutation are exactly same while the outcomes have opposite signs.

Several years ago, a reviewer for a leading informatics journal declared that "it is more than 10 years ago that anybody was interested in predicting stabilizing mutations as the problem was more or less solved." Obviously he was well too over-optimistic and our results suggest that the problem is far from solved even now as the tested algorithms are essentially no better than random assignment for new cases. We believe that the keys to the success of developing ML algorithms based methods for the purpose include the availability of significant amount of experiment data and informative features suitable for such a difficult task. While the former relies on bench scientists to perform more experiments, the later needs to be dealt with by informaticians intelligently. Useful features likely need to be based on chemical physical properties of amino acids residues around the mutation site. A possible solution is to partner ML based studies with traditional force-field based molecular simulations by deriving informative features from molecular modeling studies which can be used to model atom level interaction changes after mutations. Although force-field based simulations are demanding on computer power, recent advances in computer and software technologies allow the studies performed within a reasonable time frame.

In summary, designing protein drugs with enhanced stability using ML approaches apparently has yet reached the level for practical usages because of the limited amount of training data and unsatisfactory features. Understanding the limitations of current methods is an important step that can promote more research in this rather important field and improve research reproducibility and reliability in general.

References

1. Frokjaer S, Otzen DE (2005) Protein drug stability: a formulation challenge. Nat Rev Drug Discov 4: 298-306.
2. Pires DEV, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42: W314-W319.
3. Li Y, Zhang J, Tai D, Russell Middaugh C, Zhang Y, et al. (2012) Prots: A fragment based protein thermo-stability potential. Proteins: Structure, Function, and Bioinformatics 80: 81-92.
4. Li Y, Fang J (2012) PROTS-RF: a robust model for predicting mutation-induced protein stability changes. PLoS One 7: e47247.
5. Cheng JL, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins-Structure Function and Bioinformatics 62: 1125-1132.
6. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33: W306-310.
7. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30: 335-342.
8. Wu LC, Lee JX, Huang HD, Liu BJ, Horng JT (2009) An expert system to predict protein thermostability using decision tree. Expert Systems with Applications 36: 9007-9014.
9. Gromiha MM, Oobatake M, Sarai A (1999) Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 82: 51-67.
10. Huang LT, Gromiha MM (2009) Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. Bioinformatics 25: 2181-2187.
11. Thiltgen G, Goldstein RA (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One 7: e46084.
12. Becktel WJ, Schellman JA (1987) Protein stability curves. Biopolymers 26: 1859-1877.
13. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34: D204-206.
14. Gong H, Porter LL, Rose GD (2011) Counting peptide-water hydrogen bonds in unfolded proteins. Protein Sci 20: 417-427.