Predicting the Effect of Mutations on a Genome-wide Scale

Alexey Strokach

December 01, 2015

Contents

Abstract

[1]

[2]

[3]

1 Introduction

The computational methods for predicting protein stability upon mutation have been compared recently [**Potapov2009**].

Maximum correlation coefficient 0.86 [**Potapov2009**].

I-Mutant 2 and FoldX are trained

Rosetta can predict ddG as well, but the energy function has to be chosen carefully [**Kellogg2011**]. In particular, a soft repulsion energy function should be used for repacking (combinatorial rotamer optimization carried out using Monte Carlo simulated annealing with Dunbrack backbone dependent rotamer library), optionally combined with a hard-repulsion energy function used during backbone and sidechain minimization with uniform constraints [**Kellogg2011**]. However, optimization leads to only slightly improved accuracy for a single mutation (0.68 vs 0.69 correlation coefficient), and can be skipped in order to speed up predictions.

When the reference energies for the 20 amino acids are fit to produce the best ddG correlation with experimental values, the correlation coefficient went up to 0.73. The source code for FoldX is not availible. FoldX is trained on the Protherm and Skempi datasets, and therefore our cross-validation is likely overfitting the

2 Methods

3 Results

Untrained:

SIFT Provean MutationAssessor FATHMM (unweighted)

Trained: Polyphen-2 MutationTaster FATHMM (weighted)

4 Discussion

Depend on closed-source FoldX. Can use openMM to recreate most of the features that are used by FoldX and train a final classifier using those features.

Can use thermodynamic integration (TI) to increase the training set. Select a few mutations deemed to be the most important from each domain family. ...

5 Future Directions

References

[1] Erin D. Pleasance et al. "A comprehensive catalogue of somatic mutations from a human cancer genome". In: *Nature* 463.7278 (January 14, 2010). 00000, pp. 191–196.

[2] Jacob A. Tennessen et al. "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes". In: *Science* 337.6090 (June 7, 2012), pp. 64–69.

[3] William Lee et al. "The mutation spectrum revealed by paired genome sequences from a lung cancer patient". In: *Nature* 465.7297 (May 27, 2010), pp. 473–477.