Predicting the Effect of Mutations on a Genome-Wide Scale

by

Alexey Strokach

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Predicting the Effect of Mutations on a Genome-Wide Scale

Alexey Strokach
Master of Science
Graduate Department of Computer Science
University of Toronto
2016

Contents

Introduction

Predicting the effect of mutations is important.

1.1 Existing approaches

PoPMuSiC

Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC

mCSM: predicting the effects of mutations in proteins using graph-based signatures.

- http://www.ncbi.nlm.nih.gov/pubmed/24281696

- "To understand the roles of mutations in disease, we have evaluated their impacts not only on protein stability but also on protein-protein and protein-nucleic acid interactions".

- [1]

Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method

- http://journals.plos.org/ploscompbiol/article?id=10.1371

- "The core of the SAAMBE method is a modified molecular mechanics Poisson-Boltzmann Surface Area (MM/PBSA) method with residue specific dielectric constant".

- [2]

MAESTRO [3]

- MAESTRO implements a multi-agent machine learning system.

- Structure based tools AUTO-MUTE [7], CUPSAT [8], Dmutant [9], FoldX [10], Eris [11], PoPMuSiC [12], SDM [13] or mCSM [14] usually perform better than the sequence based counterparts. Recently, SDM and mCSM have been integrated into a new method called DUET [15].

INPS: predicting the impact of non-synonymous variations on protein stability from sequence

- http://bioinformatics.oxfordjournals.org/content/31/17/2816.long

- Here, we describe INPS, a novel approach for annotating the effect of non-synonymous mutations on the protein stability from its sequence.

- [4]

FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants

- http://journals.plos.org/ploscompbiol/article?id=10.1371

- Predict the structural effect of multiple mutations.

- "Stability effects of all possible single-point mutations were estimated using the ¡BuildModel¿ module of FoldX".

-

- We demonstrate that thermostability of the model enzymes haloalkane dehalogenase DhaA and -hexachlorocyclohexane dehydrochlorinase LinA can be substantially increased.

- [5]

1.2 Homology modeling

We used the MODELLER software package to perform all homology modeling.

1.3 Sequence features

Sorting Intolerant from Tolerant (SIFT)

The most widely-used program is SIFT. SIFT creates an extensive multiple sequence alignment for every protein, and produces a conservation score based on the likelihood of the wildtype and mutant amino acids occurring at a given position. However, we had difficulty compiling and running SIFT in a cluster setting. Furthermore, SIFT

Provean

The performance of Provean is comparable to the leading mutation scoring programs, such as SITF, PolyPhen-2, Mutation Assessor, and CONDEL [6]. Furthermore, Provean is distributed under a GPLv3 license, and uses *supporting sets* of at most 45 sequences which can be precalculated and stored. If a supporting set is available, calculating the Provean score takes several seconds per mutation.

Polymorphism Phenotyping (PolyPhen-2)

Another widely-used mutation scoring tool is PolyPhen-2. However, it is trained on a dataset of deleterious and neutral human mutations. This would make it difficult for us to run benchmarks, since we would have to be meticulous to ensure that the validation set that we are using does not contain mutations that are in the PolyPhen-2 training set.

Mutation Assessor

CONDEL

Alignments only go so far in predicting disease.

Predict loss of function much better than gain of function

1.4 Datasets

ProTherm / ProNIT [7] [8]

"MODELLER uses simulated annealing cycles along with a minimal forcefield and spatial restraints – generally Gaussian interatomic probability densities extracted from the template structure with database-derived statistics determining the distribution widthto rapidly generate candidate structures of the target sequence from the provided template sequence".

AB-Bind: Antibody binding mutational database for computational affinity predictions.

- Our Antibody-Bind (AB-Bind) database includes 1101 mutants with experimentally determined changes in binding free energies (G) across 32 complexes.

- http://www.ncbi.nlm.nih.gov/pubmed/26473627

- [9]

1.5 Benchmarks

Rosetta benchmark [10]

Benchmark showing Rosetta doing poorly: [11]

I-Mutant2, DMutant, CUPSAT, FoldX [12]

1.6 Structure features

Implementation

2.1 ELASPIC predictor

2.1.1 Training

ELASPIC described in output xxx features in total. 1. We calculated those features for the Provean and the Skempi training sets. 2. We removed features that were note different in any of the training cases (xxx for core mutations and yyy for interface mutations). 3. As described in [], balancing the training set can significantly improve performance. However, with Provean balancing the training set can bias the result because most mutations are to unconserved amino acids (often alanine) and

We built two core predictors and two interface predictors:

1. No sequence features but a balanced training set.
2. Sequence features but no balanced training set.

2.1.2 Validation

Compare how well Provean, FoldX, and 'ELASPIC with Provean' and 'ELASPIC without Provean' distinguish between the three different datasets for both core and interface mutations.

- Chaperone interaction data (core mutations) Luciferase complementation assay (interface mutations) (use Spearman correlation coefficient).
- Uniprot disease vs. polymorphism (use AUC / ROC / combination).
- COSMIC driver vs. passenger.

Chaperone interaction data

2.2 Structure features

The performance of Provean is comparable to the leading mutation scoring programs, such as SITF, PolyPhen-2, Mutation Assessor, and CONDEL [6]. Furthermore, Provean is distributed under a GPLv3 license, and uses *supporting sets* of at most 45 sequences which can precalculated and stored. If a supporting set is available, calculating the Provean score takes several seconds per mutation.

Another widely-used mutaiton scoring tool is PolyPhen-2. It is one of the packages predicted for

2.3 ELASPIC pipeline

xxx

2.4 ELASPIC web service

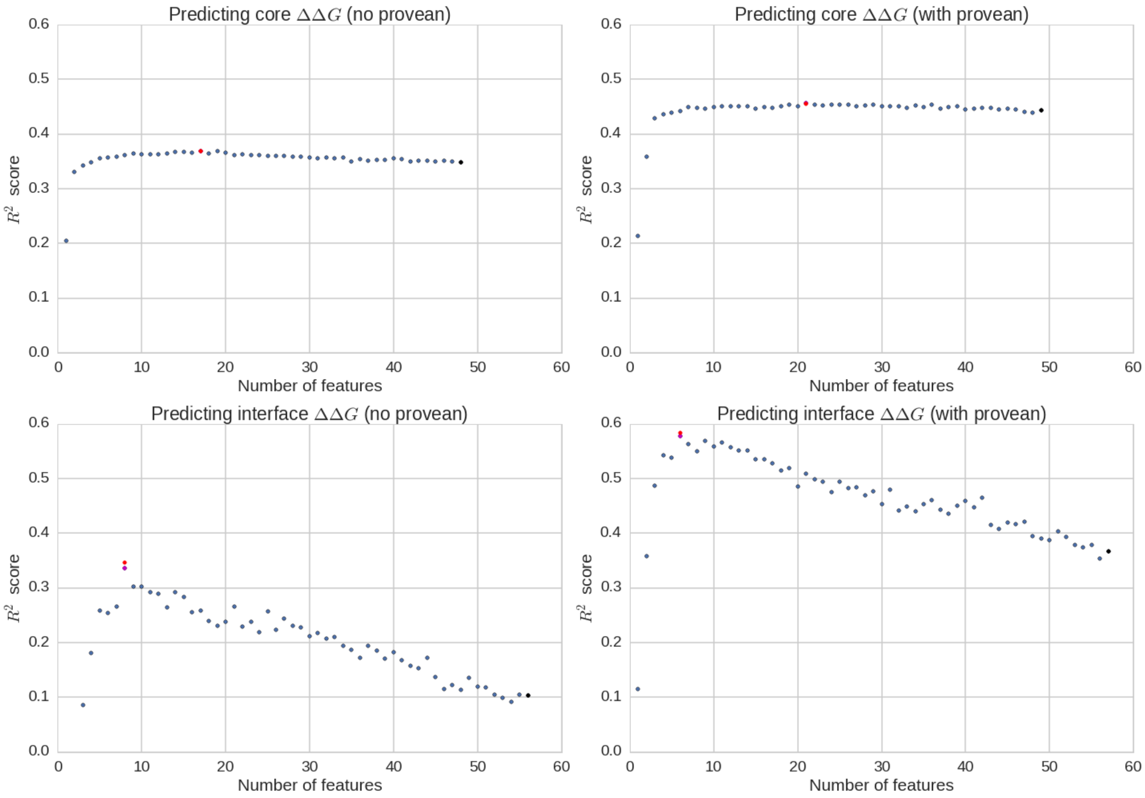


Figure 2.1: Variable elimination.

Table 2.1: ELASPIC web service API.

Method	**HTTP request**	**Description**
submitjob	POST /submitjob	Submit a job to be run on a SGE cluser.
jobstatus	GET /submitjob	View the results of a job.

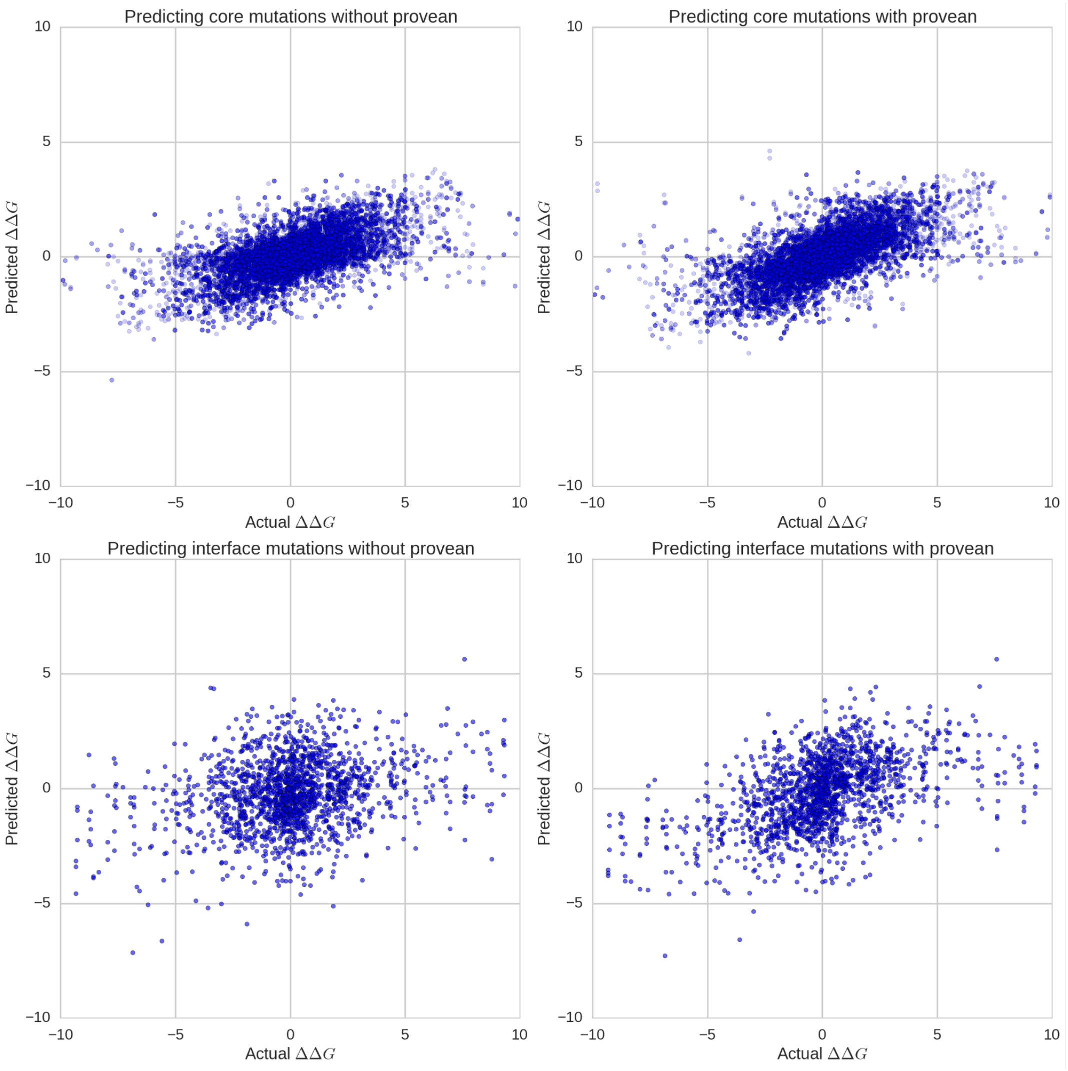


Figure 2.2: Cross-validation performance before variable elimination.

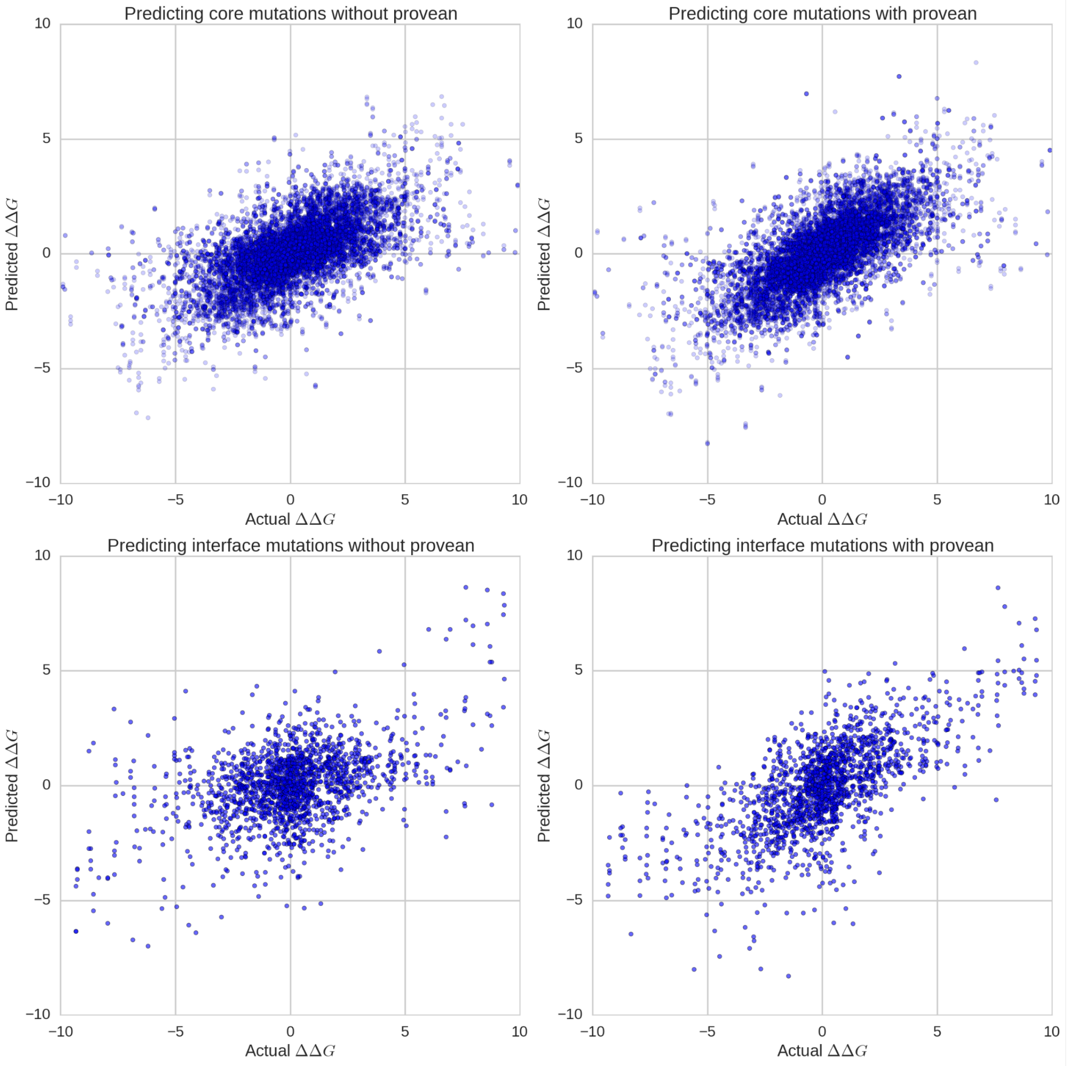
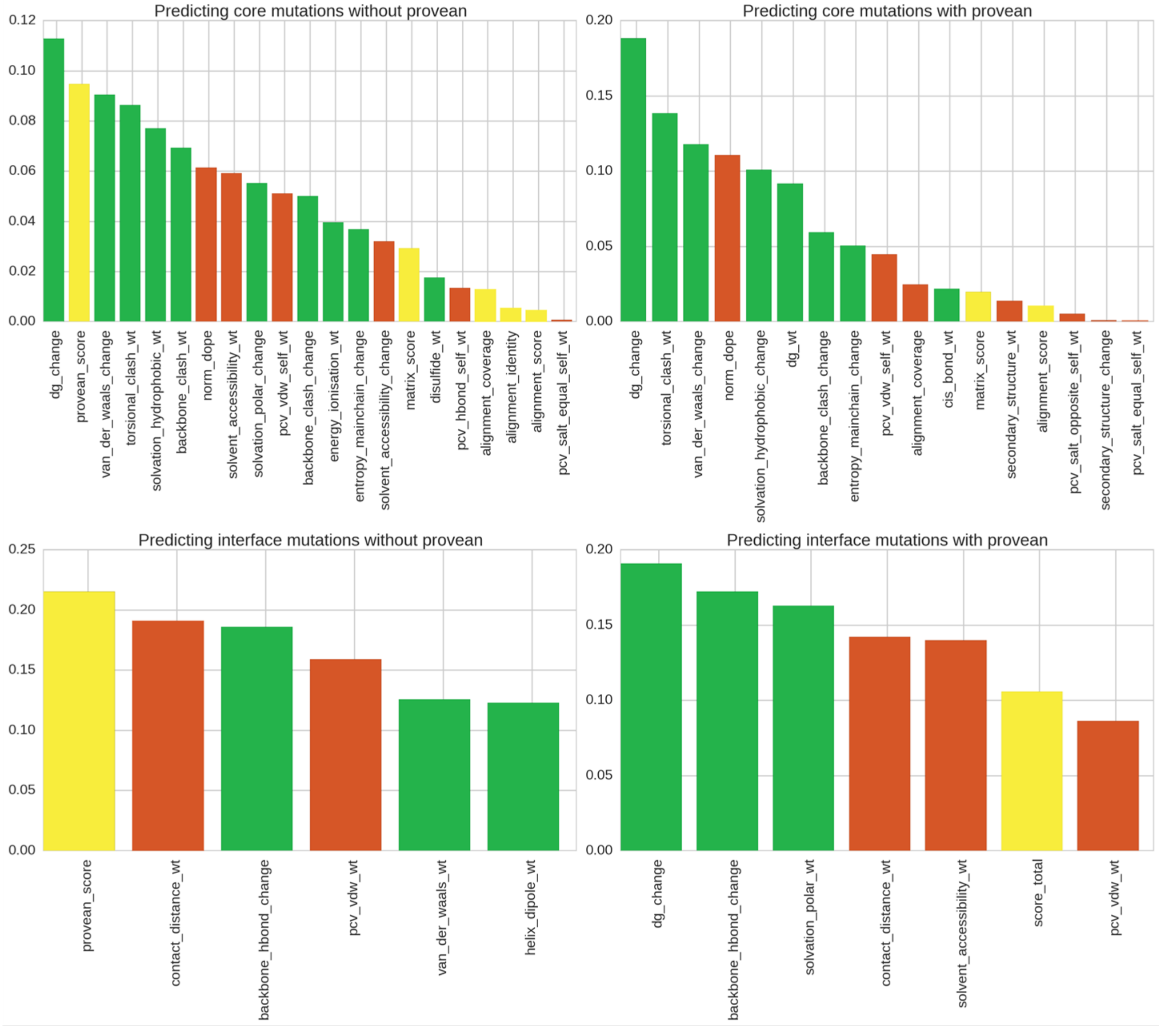


Figure 2.3: Cross-validation performance after variable elimination.



Predicting core mutations without provean

Predicting core mutations with provean

Predicting interface mutations without provean

Predicting interface mutations with provean

Figure 2.4: Feature importances after variable elimination.

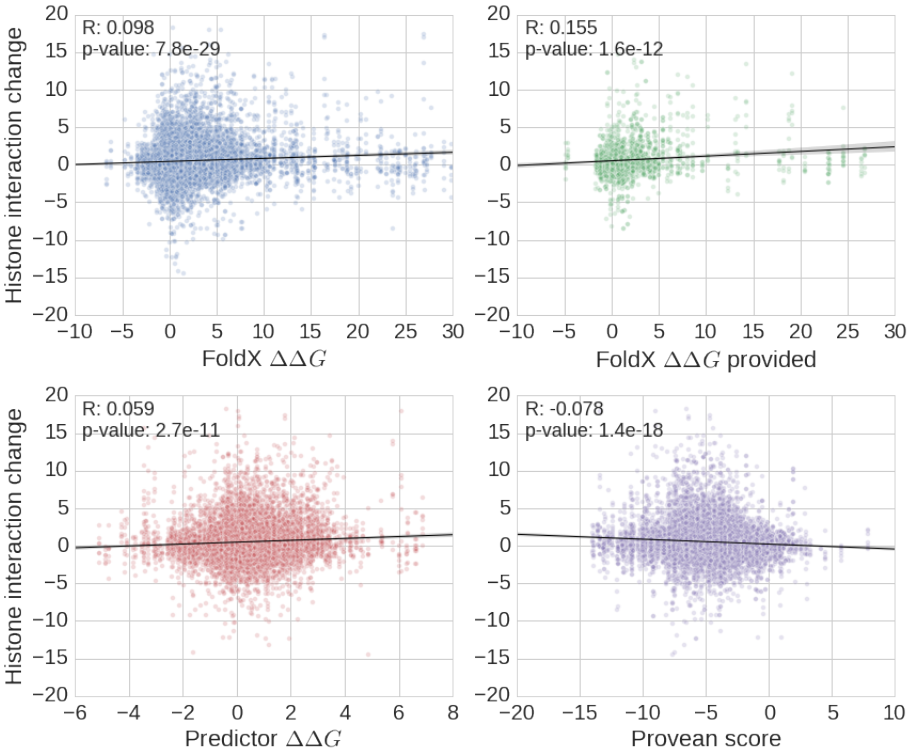


Figure 2.5: Validation of the ELASPIC core predictor using chaperone interaction data from "Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders".

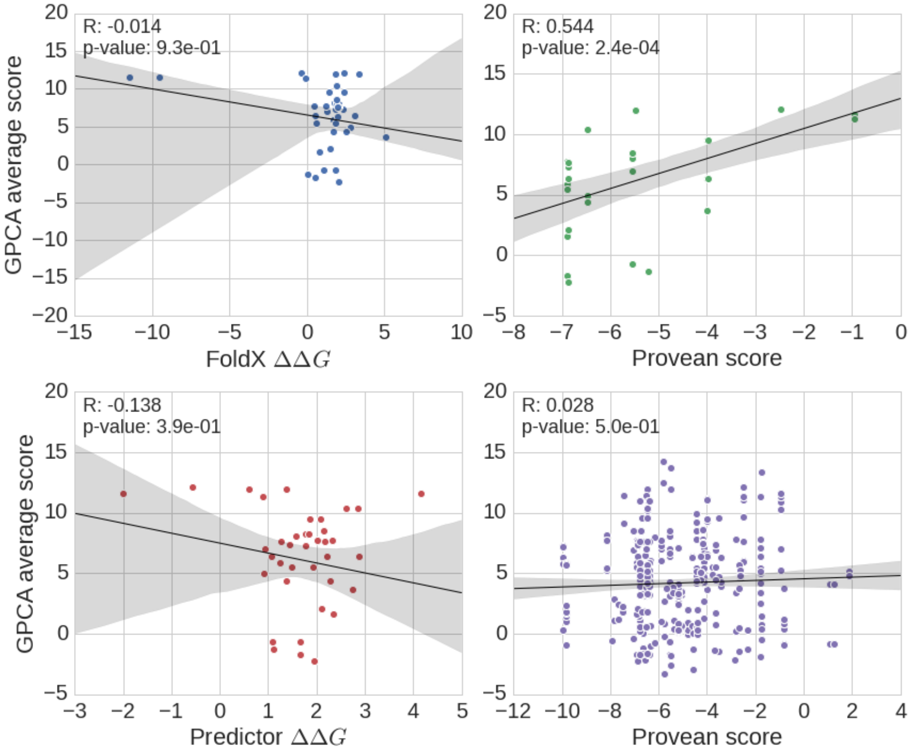


Figure 2.6: Validation of the ELASPIC core interface predictor using *Gaussia princeps* luciferase protein complementation assay from ”Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders”.

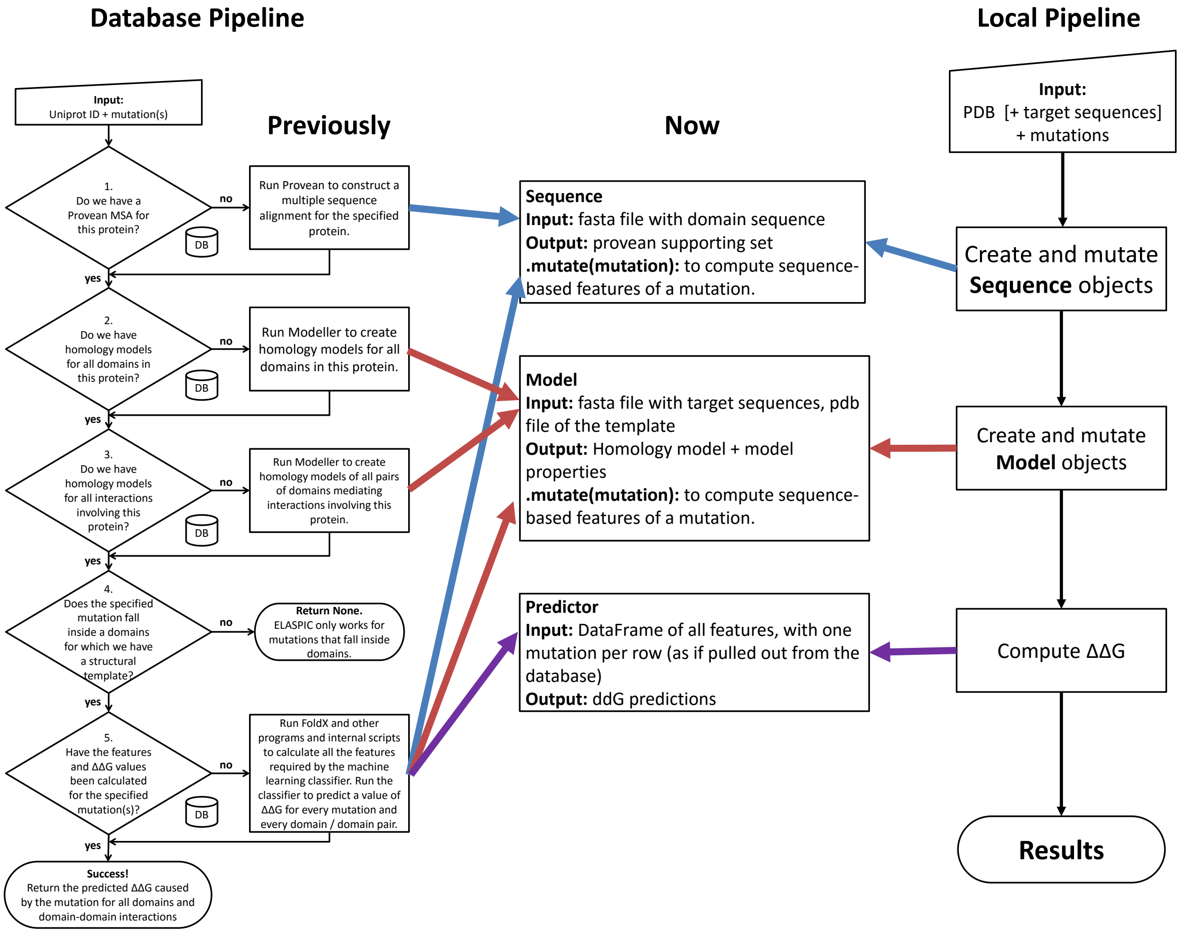


Figure 2.7: Pipeline of the ELASPIC web service.

Applications

3.1 Homology modelling of the human proteome

3.2 Energetic effect of benign and deleterious mutations

3.3 Alanine scanning of protein interfaces

- Show a histogram of $\Delta\Delta G$ values for all interfaces.

- Predict whether a peptide is going to be anti-proliferative based on the total absolute $\Delta\Delta G$ score over the interface.

3.4 Anti-proliferative peptides

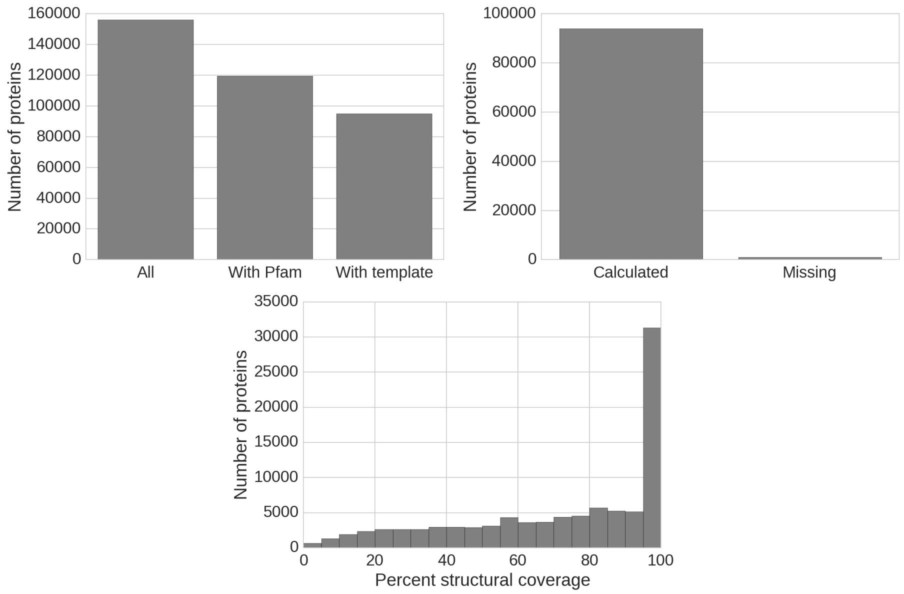


Figure 3.8: **Left**: statistics of how many homology models we were able to calculate. **Right**: structural coverage for proteins with at least one domain with a structural model.

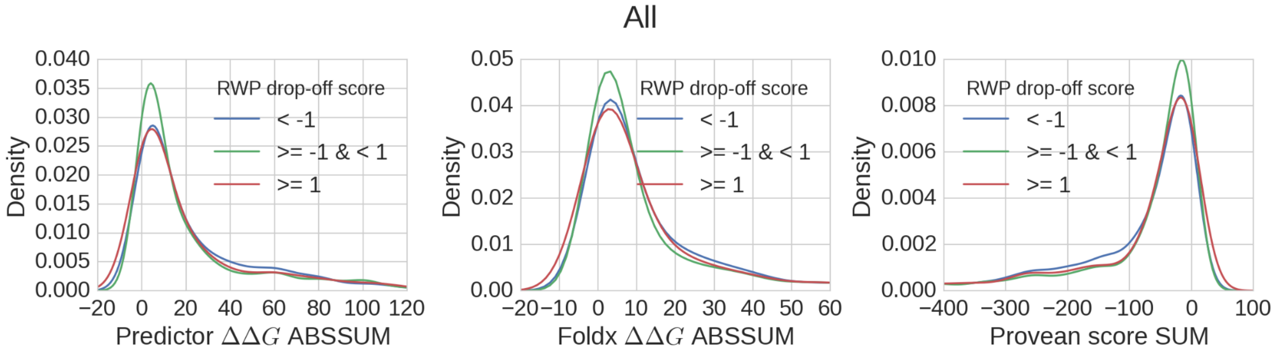


Figure 3.9: The drop-off score is higher for peptides that have a higher sum of the absolute interface FoldX energy.

Chapter 4

Discussion

4.1 Limitations

Cystic fibrosis

- Existing approaches remain limited in their ability to predict disease-causing variants. In a study of 1571 mutations of the CFTR gene causing cystic fibrosis, (SIFT, PolyPhen, PANTHER) [13]

Long QT syndrome

- Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations.
- http://www.ncbi.nlm.nih.gov/pubmed/25967940

4.2 Extensions

- Predict homo-oligomers, since this is the predominant form of oligomerization in proteins.
- Multiple amino acid substitutions + insertions / deletions
- Alternative splicing / aberrant splicing

4.3 Applications

- Predicting deleterious mutations
- Stabilizing proteins
- drugging protein-protein interfaces [14]

4.4 Domain definitions

eSCOP

Gene3D

4.5 Homology modelling

There are several approaches we could take to improve the quality of the homology models:

- Use sequence profiles (e.g. Pfam or Gene3D) to guide the alignment.
- Use multiple templates when building the homology models.
- Create multiple models and choose the one with the highest DOPE score.
- Refine the model using molecular dynamics.

Long-term MD is not useful for optimizing structures in most cases [15].

4.6 Sequence features

Improvement to sequence features

- Use covariation between amino acids in addition tho the conservation score to predict the impact of mutations, as described by Kowarsch et. al. [16].

4.6.1 Protein-DNA/RNA interactions

ProNIT

4.6.2 Protein-ligand interactions

Platinum: Protein-ligand affinity change upon mutation database.

- http://bleoberis.bioc.cam.ac.uk/platinum/

BioLiP is a semi-manually curated database for high-quality, biologically relevant ligand-protein binding interactions.

- http://zhanglab.ccmb.med.umich.edu/BioLiP/

- The structure data are collected primarily from the Protein Data Bank, with biological insights mined from literature and other specific databases.

4.6.3 Predicting PPI

PRISM: Protein interaction by structure matching

Bibliography

[1] Douglas E. V. Pires et al. “mCSM: predicting the effects of mutations in proteins using graph-based signatures”. In: *Bioinformatics* 30.3 (January 2, 2014), pp. 335–342.

[2] Marharyta Petukh et al. “Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method”. In: *PLOS Comput Biol* 11.7 (July 6, 2015), e1004276.

[3] Josef Laimer et al. “MAESTRO - multi agent stability prediction upon point mutations”. In: *BMC Bioinformatics* 16 (2015), p. 116.

[4] Piero Fariselli et al. “INPS: predicting the impact of non-synonymous variations on protein stability from sequence”. In: *Bioinformatics* 31.17 (January 9, 2015), pp. 2816–2821.

[5] David Bednar et al. “FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants”. In: *PLOS Comput Biol* 11.11 (November 3, 2015), e1004556.

[6] Yongwook Choi et al. “Predicting the Functional Effect of Amino Acid Substitutions and Indels”. In: *PLoS ONE* 7.10 (October 8, 2012). 00256, e46688.

[7] Jianyang Zeng et al. “A Bayesian approach for determining protein side-chain rotamer conformations using unassigned NOE data.” In: *Journal of computational biology : a journal of computational molecular cell biology* 18.11 (2011), pp. 1661–79.

[8] M. D. Shaji Kumar et al. “ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions”. In: *Nucleic Acids Research* 34 (suppl 1 January 1, 2006), pp. D204–D206.

[9] Sarah Sirin et al. “AB-Bind: Antibody binding mutational database for computational affinity predictions”. In: *Protein Science* 25.2 (February 1, 2016), pp. 393–409.

[10] Shane Ó Conchúir et al. “A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design”. In: *PloS One* 10.9 (2015), e0130433.

[11] Vladimir Potapov et al. “Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details”. In: *Protein Engineering Design and Selection* 22.9 (January 9, 2009), pp. 553–560.

[12] Sofia Khan and Mauno Vihinen. “Performance of protein stability predictors”. In: *Human Mutation* 31.6 (June 1, 2010), pp. 675–684.

[13] R Dorfman et al. “Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene?” In: *Clinical Genetics* 77.5 (May 1, 2010), pp. 464–473.

[14] James A. Wells and Christopher L. McClendon. “Reaching for high-hanging fruit in drug discovery at protein–protein interfaces”. In: *Nature* 450.7172 (December 13, 2007), pp. 1001–1009.

[15] Alpan Raval et al. “Refinement of protein structure homology models via long, all-atom molecular dynamics simulations”. In: *Proteins: Structure, Function, and Bioinformatics* 80.8 (August 1, 2012), pp. 2071–2079.

[16] Andreas Kowarsch et al. “Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions”. In: *PLoS Comput Biol* 6.9 (September 16, 2010), e1000923.