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

List of Tables

List of Figures

Implementation

2.1 Training sets

2.1.1 Core

Sanhi et al. databaset (taipale)

2.1.2 Interface

2.1.3 Standalone pipeline

2.1 right

2.2 ELASPIC pipeline

An overview of the ELASPIC pipeline is presented in Figure 2.1. ELASPIC includes a library Python scripts for construction sequence alignments, constructing Provean supporting sets and computing the Provean score, constructing homology models, running FoldX, and predicting the $\Delta\Delta G$ of the mutation. It also includes a "Standalone Pipeline" and a "Database Pipeline", which include command line options for mutating a protein structure.

2.2.1 Standalone pipeline

The standalone pipeline works without downloading and installing a local copy of the ELASPIC and PDB databases, but requires a PDB structure or template to be provided for every protein. Pipeline output is saves as JSON files inside the working directory, rather than being uploaded to the database as in the case of the database pipeline. The general overview of the local pipleine is presented in the figure to the right.

The local pipeline still requires a local copy of the Blast nr database.

2.2.2 Database pipeline

The database pipeline allows mutations to be performed on a proteome-wide scale, without having to specify a structural template for each protein. This pipeline requires a local copy of ELASPIC domain definitions and templates, as well as a local copy of the BLAST and PDB databases.

The general overview of the database pipleine is presented in 2.1 left. A user runs the ELASPIC pipeline specifying the Uniprot ID of the protein being mutated, and one or more mutations affecting that

protein. At each decision node, the pipeline queries the database to check whether or not the required information has been previously calculated. If the required data has not been calculated, the pipeline calculates it on the fly and stores the results in the database for later retrieval. The pipeline proceeds until homology models of all domains in the protein, and all domain-domain interactions involving the protein, have been calculated, and the $\Delta\Delta G$ has been predicted for every specified mutation.

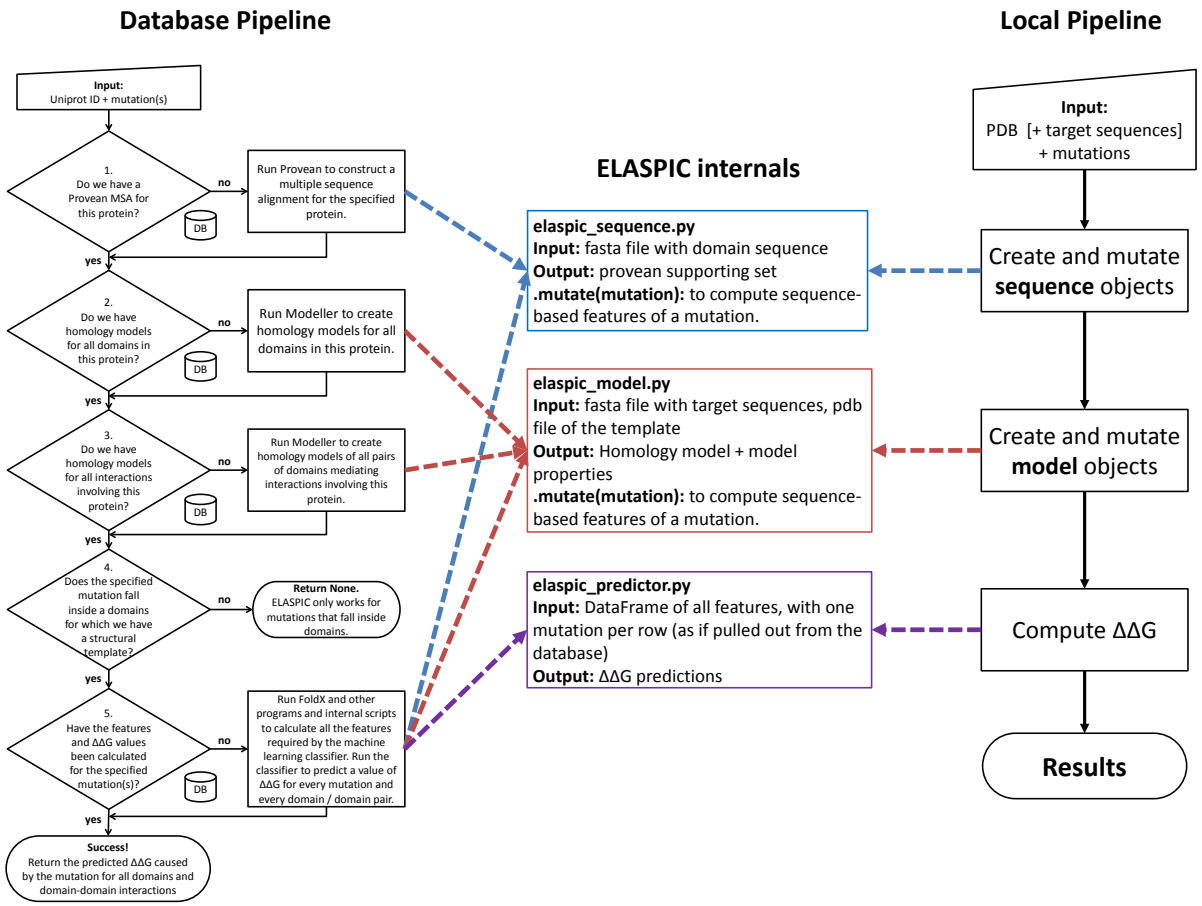

Figure 2.1: Overview of the ELASPIC pipeline. A user runs the ELASPIC pipeline specifying the UniProt id of the protein being mutated, and one or more mutations affecting that protein. At each decision node, the pipeline queries the database to check whether or not the required information has been calculated previously. If the required data has not been calculated, the pipeline calculates it on the fly and stores the results in the database for later retrieval. The pipeline proceeds until homology models of all domains in the protein, and all domain-domain interactions involving the protein, have been calculated, and the $\Delta\Delta G$ has been predicted for every specified mutation.

2.3 ELASPIC predictor

ELASPIC uses the gradient boosting of decision trees regressor (GBR). It was optimized in several ways.

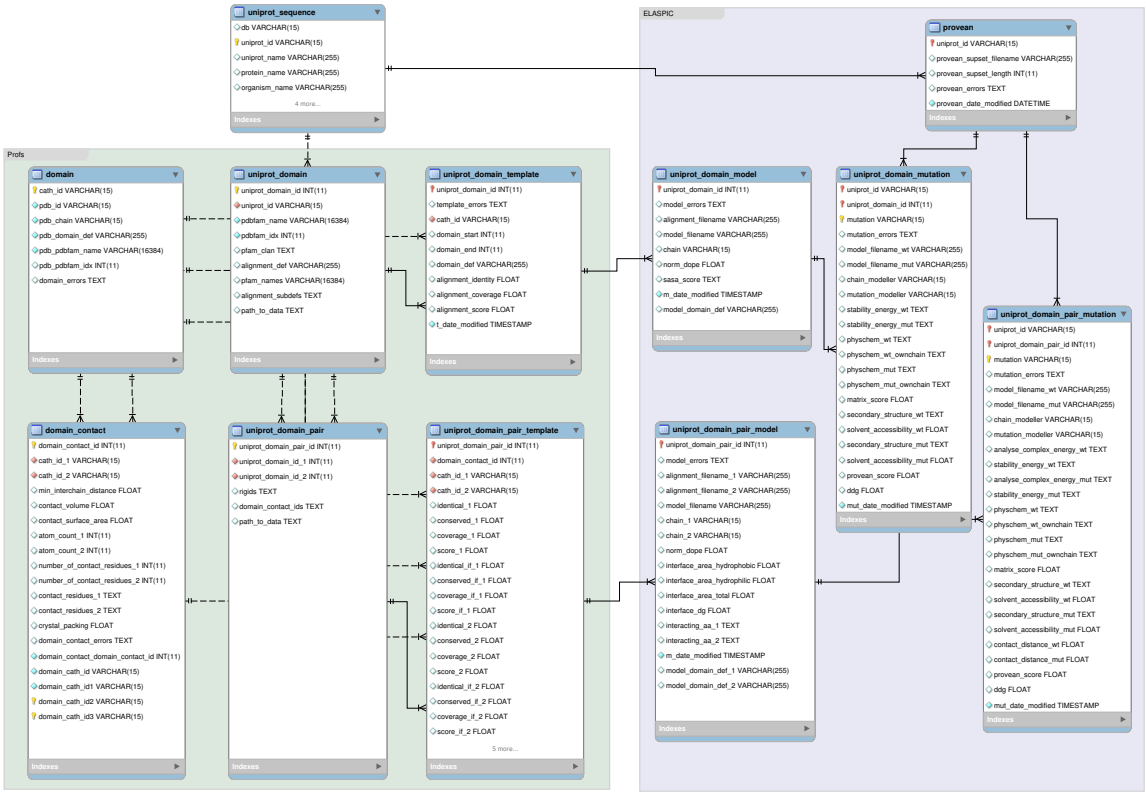

Figure 2.2: Database schema used by the ELASPIC pipeline. Tables on the green plate titled Profs are calculated using the Profs pipeline, as described in [1]. Tables on the purple plate titled ELASPIC are calculated using the ELASPIC pipeline, following the procedure outlined in 2.1. A detailed description of each table can be found in **??**.

2.3.1 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. It has been reported that balancing the training set by including both positive and negative samples

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.

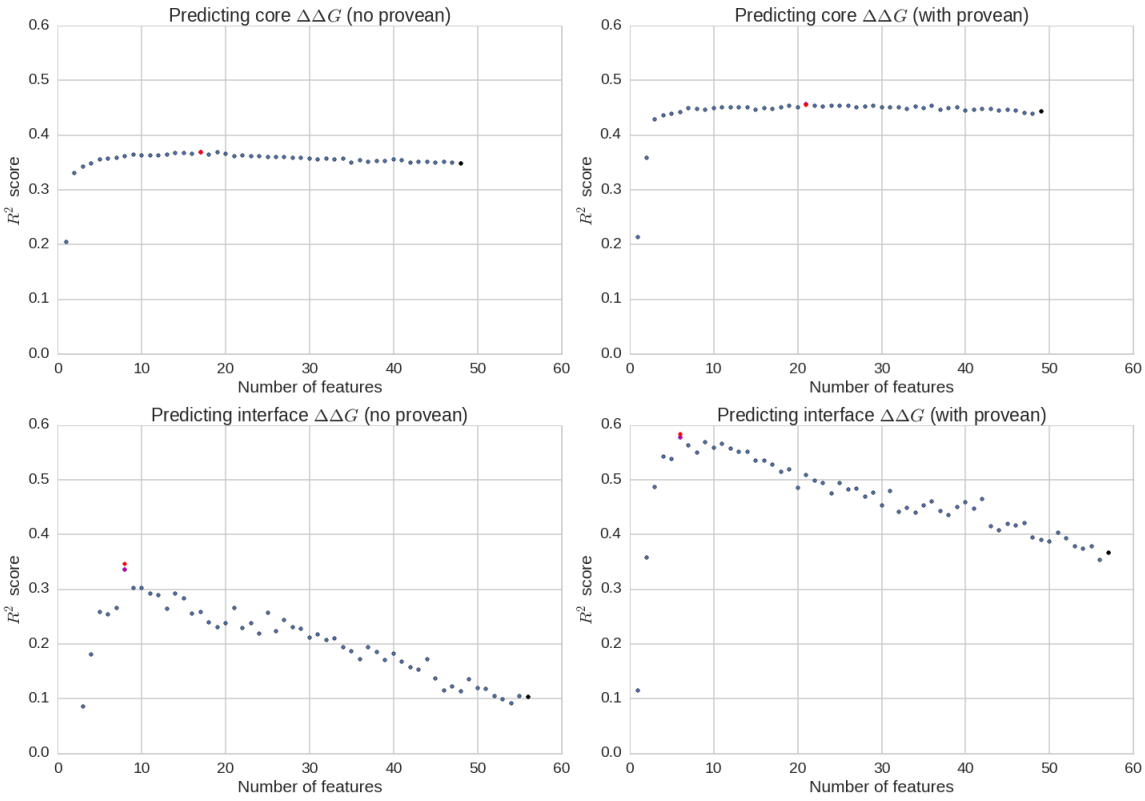

Figure 2.3: Variable elimination.

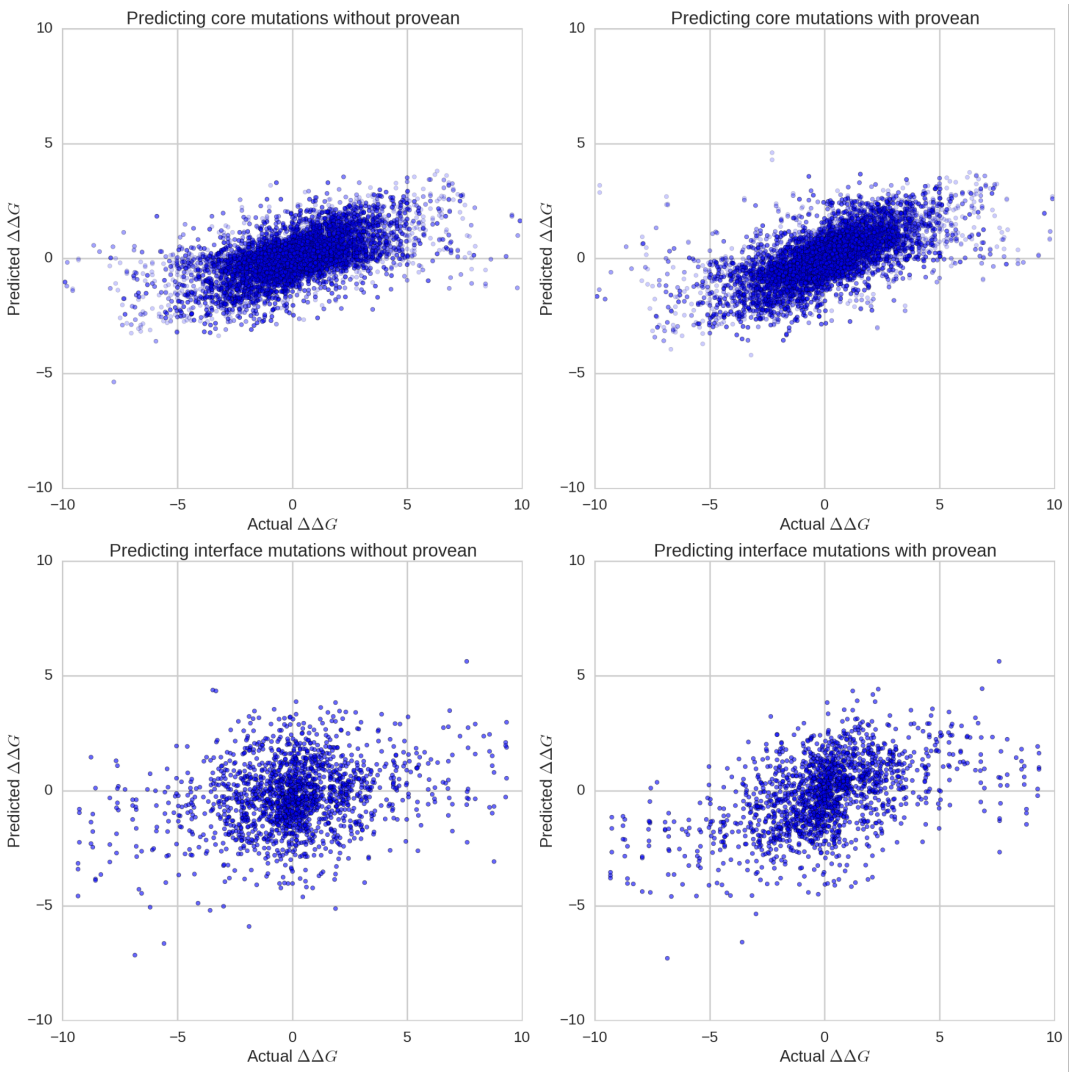

Figure 2.4: Cross-validation performance before variable elimination.

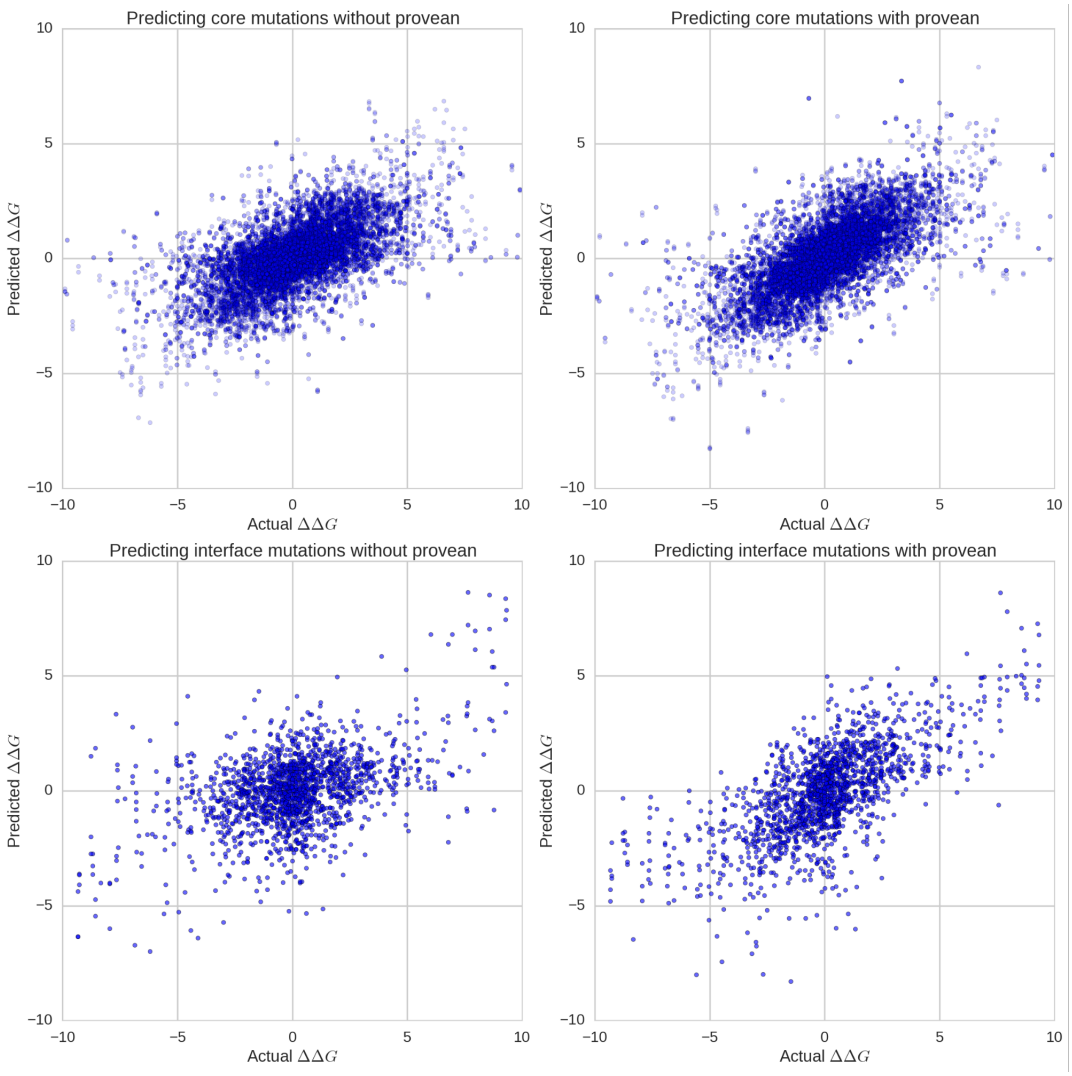

Figure 2.5: Cross-validation performance after variable elimination.

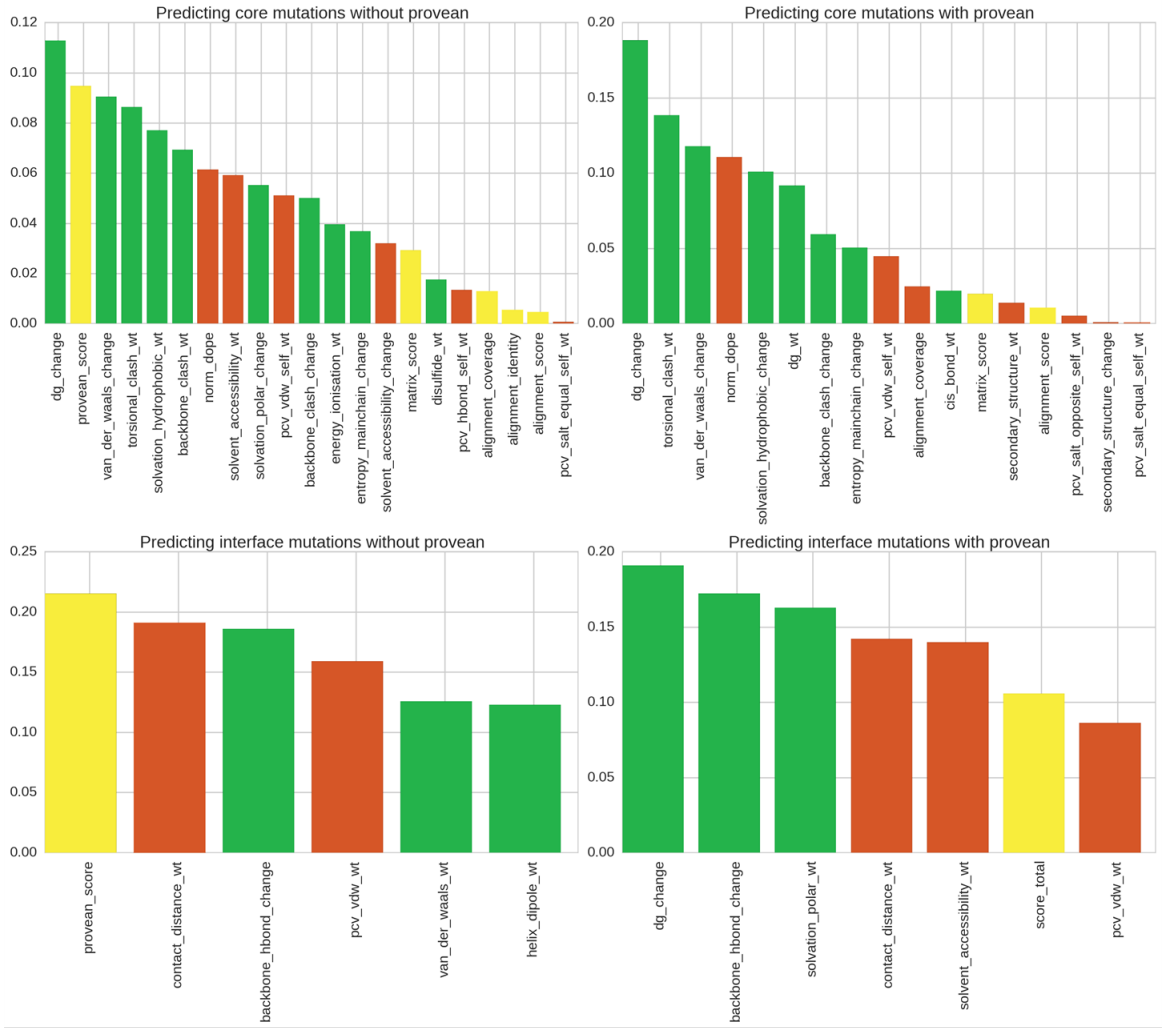

Figure 2.6: Feature importances after variable elimination.

2.3.2 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.

2.4 Structure features

The performance of Provean is comparable to the leading mutation scoring programs, such as SITF, PolyPhen-2, Mutation Assessor, and CONDEL [3]. 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

2.5 ELASPIC pipeline

The ELASPIC project was started by Niklas Berliner and others in 2014 [7].

ELASPIC uses Modeller [8] to construct homology models of domains and domain-domain interactions, FoldX to optimize those model and to introduce mutations [9], and the ELASPIC predictor to combine FoldX energy scores with sequence-based and other features and predict the energetic impact of a mutation on the stability of a single domain or the affinity between two domains. A flowchart describing the ELASPIC pipeline is presented in 2.1. At each step in the pipeline, a local database is queried to see if the required information has already been calculated. If the information is available, the pipeline moves to the next step. If the information is not available, the pipeline runs the module that generates the required information, stores the generated information in the database for future access, and then moves to the next step. If the specified mutation falls outside of every domain in the protein, no predictions are returned. Otherwise, the pipeline evaluates the impact of the mutation on the stability of the domain and, if the mutation falls in a domain interface, on the affinity between two domains. In order to expedite the evaluation of mutations, we precalculated homology models and Provean supporting sets for all human proteins. Structural and sequential features, and predicted G scores, have also been precalculated for the majority of mutations listed in the Uniprot humsavar file [10] and in the COSMIC [11] and ClinVar [12] databases.

Provean supporting sets, homology models and mutation G scores are available from the ELASPIC downloads page: http://elaspic.kimlab.org/static/download/. The source code of the python package implementing the ELASPIC pipeline is available from https://github.com/kimlaborg/elaspic, and the documentation for the ELASPIC pipeline can be accessed online at http://elaspic.readthedocs.org/.

2.6 ELASPIC web service

Table 2.2: 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.

Table 2.1: ELASPIC database tables.

Table name	Table description
domain	Contains Profs domain definitions for all proteins in the PDB.
domain_contact	Contains information about interactions between Profs domains in the PDB. Only interactions that are predicted to be real by NOXclass [2] are included in this table.
uniprot_sequence	Contains protein sequences for all proteins that are annotated with Profs domains in the **uniprot_domain** table. This table is constructed by downloading and parsing *uniprot_sprot_fasta.gz*, *uniprot_trembl_fasta.gz*, and *homo_sapiens_variation.txt* files from the Uniprot.
provean	Contains information about Provean [3] supporting set files. The construction of a supporting set is the longest part of running Provean. Thus, in order to speed up the evaluation of mutations, the supporting set is precalculated and stored for every protein.
uniprot_domain	Contains Profs domain definitions for proteins in the **uniprot_sequence** table. This table is obtained by downloading Pfam domain definitions for all known proteins from SIMAP [4], and mapping those proteins to Uniprot using the MD5 hash of each sequence. Overlapping and repeating domains are either merged or deleted, as described in [1].
uniprot_domain_template	Contains structural templates for domains in the **uniprot_domain** table. The *domain_def* column contains expanded and corrected domain definitions for every domain.
uniprot_domain_model	Contains information about the homology models which were created using structural templates in the **uniprot_domain_template** table.
uniprot_domain_mutation	Contains information about the structural impact of core mutations, calculated by introducing those mutations into homology models listed in the **uniprot_domain_model** table. The *ddg* column contains the predicted change in the Gibbs free energy of binding.
uniprot_domain_pair	Contains pairs of domains that are likely to mediate the interaction between known interacting partners, obtained from Hippie [5] and Rolland et al. [6].
uniprot_domain_pair_template	Contains structural templates for domain pairs in the **uniprot_domain_pair** table.
uniprot_domain_pair_model	Contains information about homology models which were created using structural templates in the **uniprot_domain_pair** table.
uniprot_domain_pair	Contains information about the structural impact of interface mutations, calculated by introducing those mutations into homology models listed in the **uniprot_domain_pair_model** table. The *ddg* column contains the predicted change in the Gibbs free energy of binding.

Bibliography

[1] Daniel K. Witvliet et al. "ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity". In: *Bioinformatics* 32.10 (May 15, 2016), pp. 1589–1591. DOI: 10.1093/bioinformatics/btw031.

[2] Hongbo Zhu et al. "NOXclass: prediction of protein-protein interaction types". In: *BMC Bioinformatics* 7.1 (January 19, 2006), p. 27. DOI: 10.1186/1471-2105-7-27.

[3] Yongwook Choi et al. "Predicting the Functional Effect of Amino Acid Substitutions and Indels". In: *PLoS ONE* 7.10 (October 8, 2012). 00256, e46688. DOI: 10.1371/journal.pone.0046688.

[4] Thomas Rattei et al. "SIMAP—a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters". In: *Nucleic Acids Research* 38 (suppl 1 January 1, 2010). 00031, pp. D223–D226. DOI: 10.1093/nar/gkp949.

[5] Martin H. Schaefer et al. "HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores". In: *PLoS ONE* 7.2 (February 14, 2012), e31826. DOI: 10.1371/journal.pone.0031826.

[6] Thomas Rolland et al. "A Proteome-Scale Map of the Human Interactome Network". In: *Cell* 159.5 (November 20, 2014). 00006, pp. 1212–1226. DOI: 10.1016/j.cell.2014.10.050.

[7] Niklas Berliner et al. "Combining Structural Modeling with Ensemble Machine Learning to Accurately Predict Protein Fold Stability and Binding Affinity Effects upon Mutation". In: *PLoS ONE* 9.9 (September 22, 2014), e107353. DOI: 10.1371/journal.pone.0107353.

[8] Benjamin Webb and Andrej Sali. "Comparative Protein Structure Modeling Using MODELLER". In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002.

[9] Joost Schymkowitz et al. "The FoldX web server: an online force field". In: *Nucleic Acids Research* 33 (suppl 2 January 7, 2005), W382–W388. DOI: 10.1093/nar/gki387.

[10] The UniProt Consortium. "UniProt: a hub for protein information". In: *Nucleic Acids Research* 43 (D1 January 28, 2015), pp. D204–D212. DOI: 10.1093/nar/gku989.

[11] Simon A. Forbes et al. "COSMIC: exploring the world's knowledge of somatic mutations in human cancer". In: *Nucleic Acids Research* 43 (D1 January 28, 2015), pp. D805–D811. DOI: 10.1093/nar/gku1075.

[12] Melissa J. Landrum et al. "ClinVar: public archive of interpretations of clinically relevant variants". In: *Nucleic Acids Research* 44 (D1 April 1, 2016), pp. D862–D868. DOI: 10.1093/nar/gkv1222.