ACDC-NN

ACDC-NN is a novel antisymmetric neural network to predict proteins free energy changes upon point variations along the amino acid sequence.
The ACDC-NN model was built so that it can be used to make predictions in two different ways:

1. when both the wild-type and variant structure are available, these are respectively used as direct and inverse inputs so that the network can provide a prediction that, by construction, is perfectly antisymmetric;

2. when only the wild-type structure is available, as usual, the input for the inverse substitution is created starting from the direct one by inverting the variation encoding but preserving the same structure.

For further information about the ACDC-NN architecture and properties, please see the related paper https://doi.org/10.1088/1361-6463/abedfb

ACDC-NN Seq is a sequence-based version of ACDC-NN that does not require the structure of the protein, further information is available in the paper:
https://doi.org/10.3390/genes12060911

About this repository

Here you can find the instructions to easily install ACDC-NN on your computer using pip (see commands below).
In this version, ACDC-NN was trained using all datasets available in the literature without correcting for sequence similarity.
In case you want to replicate our paper results you will find a jupyter notebook inside the 'results_replication' folder.
There ACDC-NN was trained using a 10-fold cross-validation taking into account sequence similarity to avoid overfitting.

Installation

We recommend using pip:
```
pip install acdc-nn
```

Requirements:
<table>
  <tr><th>Requirement</th><th>Minimum tested version</th></tr>
  <tr><td>python</td><td>3.6</td></tr>

  <tr><td>tensorflow</td><td>2.3.1</td></tr>
  <tr><td>Biopython</td><td>1.78</td></tr>
  <tr><td>numpy</td><td>1.19.5</td></tr>
  <tr><td>pandas</td><td>1.1.5</td></tr>
  <tr><td>silence_tensorflow</td><td>1.1.1</td></tr>
</table>

Usage

To predict the change of the folding free energy (DDG) due to a point substitution in a protein sequence, ACDC-NN needs both evolutionary and structural information about the protein itself. The structural information is from a PDB file. The evolutionary information is from a profile file, simple tab-separated table of the frequencies of each residue in each position in homologous proteins. Positive DDG values are stabilizing.

When no structural information is available, the sequence-based ACDC-NN Seq network must be used:
```
acdc-nn seq SUB PROFILE

When information is available only for the wild-type protein, the predictor can be run as:
```
acdc-nn struct SUB PROFILE PDB CHAIN
```
where SUB is the point substitution, PROFILE and PDB are the paths to the profile and PDB files, and CHAIN is the PDB chain where the substitution occurs. SUB is in the form XNY where X is the wild-type residue, N is the position of the substitution, and Y is the mutated residue. X and Y are given as a one-letter amino acid code and N is 1-based and referred to the PDB numbering of the relevant chain, and not the position in the sequence. Both PDB and profile files are automatically decompressed when they have a ".gz" extension.

When information is available also for the mutated protein, a better prediction can be got as:
```
acdc-nn istruct SUB WT-PROFILE WT-PDB WT-CHAIN INV-SUB MT-PROFILE MT-PDB MT-CHAIN
```

To predict more than a few substitutions, we provide a batch mode:
```
acdc-nn batch SUBS
```
where SUBS is the path to a tab-separated table with a row for each substitution to be predicted.
```

For substitutuion where no structural information is available the row format is:

```
SUB PROFILE
```

For substitutions where only the wild-type protein data is available, the row format is:

```
SUB PROFILE PDB CHAIN
```

For substitutions where also the mutated protein data is available, the row format is:

```
SUB WT-PROFILE WT-PDB WT-CHAIN INV-SUB MT-PROFILE MT-PDB MT-CHAIN
```

The three formats can be mixed arbitrarily in the same file.

Examples

These examples use the data in the tests directory of the github repository.
No structure available:

```
> acdc-nn seq Q104H tests/profiles/2ocjA.prof.gz

0.06451824
```

Single substitution:

```
> acdc-nn struct Q104H tests/profiles/2ocjA.prof.gz tests/structures/2ocj.pdb.gz A

0.15008962
```

Single substitution with the structure of the mutated protein

```
> acdc-nn istruct V51I tests/profiles/1bsaA.prof.gz tests/structures/1bsa.pdb.gz A I51V tests/profiles/1bniA.prof.gz tests/structures/1bni.pdb.gz A

0.48577148

> acdc-nn istruct I51V tests/profiles/1bniA.prof.gz tests/structures/1bni.pdb.gz A V51I tests/profiles/1bsaA.prof.gz tests/structures/1bsa.pdb.gz A

-0.48577148
```

NB: In the above example we have specifically chosen two homologous proteins that have similar structure.