Syllabus for protein structure prediction hands-on

This project is a follow-up of the Salmonella's SNP project. The goal is to model the structure of the protein that carries the mutation responsible for high resistance to tetracyclin of a Salmonella variant.

Outline of your project

Predict a protein structure

You will model the protein structure of the provided sequence (see data/ folder on the Git).

For this you need to fetch the multiple sequence alignment that corresponds to this protein family.

Then, you will develop a piece of software that detects the covariations from the sequences. You will transform this information into a contact map (a matrix that tells which amino-acid are in contact in the 3D structure).

Then use the modeller FT-COMAR in the tools/ directory to create the 3D structure.

Check your structure

Once you have generated the 3D structure, compare it to the real structure by finding the reference on the biggest database of publicly available protein structures Protein Data Bank.

Visualize (using PyMol for instance) and align the model to the template you have downloaded.

Quantify the similarity between the model and the reference.

Optimize your predictions

There are several ways of improving your predictions. Read the suggestions in the Extension section.

Project submission and evaluation

You will work in team of 2-3 people. You'll submit your code and a 3-5 pages report describing results, tests and optimization of your method.

Clarity, trustworthiness of the tests and method will be the most important criteria for the evaluation.

Guidelines

Fetch a multiple sequence alignment (MSA)

The provided sequence belong to a specific protein family (meaning that many homologous sequences are known in various organisms). Use the databases and tools detailed in the section below to find the associated protein family, and fetch a multiple sequence alignment for this family.

There are multiple choice in the size of the MSA, depending on the degree of redundancy reduction. Choose a size such that you have at least 1000 sequences in order to have an accurate estimation of residues covariations.

Predicting contact map from multiple sequence alignment (MSA)

There are 20 standard amino acids. They have various physicochemical properties, such as size, charge, aromatic cycle, etc.

The structure and therefore the interactions between the amino acid of the protein sequence determine its function in the organism. Facing amino-acid in 3D structure interacts, we say there are in *contact*. The structure of the protein is stable mainly due to electrostatic interaction between amino-acid being in contact.

Therefore if one mutates, the amino-acid in contact have to compensate the mutation: e.g. if two amino-acid X and Y are charged positively and negatively respectively, if X is mutated in a negatively charged amino-acid, then Y have to be mutated in a positively charged amino-acid.

By aligning mulitple related protein sequences and computing the mutual information between the position of the alignment, one can infer the possible contacts.

Some notations

$s_{i,p}$ represents the amino-acid at position p in the i$^{\text{th}}$ sequence. X_p is the random variable representing the amino-acid at position p.

How to quantify the covariations?

You want to know if the amino-acid at a position p determines the amino-acid at position q, and vice versa.

One simple way to quantify this is to use the mutual information between X_p and X_q, which we will denote by $MI_{p,q}$.

Make sure you fully understand by answering the following questions:

- if a position p is fully conserved across sequences, what will $MI_{p,q}$ be equal to?
- if a position p is independent from a position q, what will $MI_{p,q}$ be equal to?
- when a couple of positions is expected to have a maximum mutual information?
- plot the entropy of each position of the alignment, as well as the heatmap of the mutual information. What do you observe?

Suppose that a group of proteins in your MSA comes from a relatively recent common ancestor. How is it impacting your prediction? Can you find a way to correct for this?

Predicting the contact map

Develop a module that creates a contact map (see file format below to be compatible with FT-COMAR) when providing a MSA as input.

From CM to structure

FT-COMAR is a simple tool to predict the structure of a protein based on its contact map. You can find it in the Git.

Extensions

In these extensions, we propose to optimize and learn models to replace the manually set parameters. We will rely on a fully automated pipeline to predict the structure and compute its RMSD with the reference structure.

Optimize the thresholding

One crucial parameter is the threshold on the mutual information that decides for a contact between residues.

Choose between the following possibilities: - Create a full pipeline that, from an MSA and a target structure, computes the RMSD between the the predicted structure and the target structure. - Compute the expected CM of the target and optimize the threshold on MI to match as closely as possible the target CM.

What are the (dis)advantages of the two approaches?

Learn the sequence weights

Develop a model that takes the MSA (or statistics computed on it) as input and output the sequence weights used to predict the structure. The model should be learn so that the predicted structure is as close as possible as the final structure.

Find a biological story

Check where the mutation occurs in the protein structure and propose a scenario that would explain the higher resistance to the tetracyclin antibiotic.

Tips

Read files in standard format (FASTA)

Most of the sequence files in bioinformatics are in a very simple format called FASTA.

The extensions are usually:

Sequence_type	Extension
Nucleic	.fna
Amino acid	.faa
Unknown	.fas

The submodule SeqIO of Biopython is able to read fasta files as the following code shows:

```
import Bio.SeqIO

for i,record in enumerate(Bio.SeqIO.parse("path/to/my/file.fasta","fasta"):
        if i%1000==0:
            print(i)
        print(str(record.record),":",str(record.seq))
```

Databases and on-line tools

Pfam

Pfam is a database of protein families (i.e. protein that share a common ancestor and that are supposed to play similar roles in the organisms they are expressed in).

You can retrieve pre-computed multiple sequence alignments of known protein families (in the **Alignments** tab in the menu on the left). These protein sequence alignments can be used for many purposes, and in particular for protein structure prediction (see following section).

Protein structure prediction

FT-COMAR

This software generates a 3D structure from a contact map.

You can download it here: FT-COMAR.

The contact map have to be in the following format:

```
133
1
```

```
0
0
0
0
1
0
...
```

Where the first line is the number N of amino-acid in the sequence, and the following N^2 lines indicate if there is a **contact** by a 1 between amino acid i and amino acid j, or **no contact by a 0**. The order of the lines correspond to the lexicographic order: (1,1), (1,2),. . . , (1,N), (2,1), (2,2),. . . , (2,N), . . . , (N,N).

Here is an example for getting the 3D structure (in PDB format) associated to the contacts listed in `my_contacts.lst`:

```
path/to/FT-COMAR my_contacts.lst 9 0 test.pdb
```

If you want to visualize the obtained 3D structure, you can use rasmol by calling:

```
pymol test.pdb
```

PyMol

To align protein structures, just load the two structures - say predicted_prot and known_prot - in PyMol and run the command:

```
cealign predicted_prot,known_prot
```

This will optimize the RMSD of the alignment using the CE algorithm.