Laboratory of Systems and Synthetic Biology
Wageningen University and Research Center
Dreijenplein 10, Wageningen
6703 HB, the Netherlands

Maarten Reijnders
E: maarten.reijnders@wur.nl
Building 316

Thesis project:

Consensus-based protein function prediction

Keywords: Bioinformatics, protein annotation, algae

The revolution in sequencing technology is resulting in a multitude of genomes being sequenced. For many genes in such newly obtained genomes, it is unclear what their function is. Homology-based functional annotation methods are capable of annotating large proportions of these genomes by using pre-annotated genomes as a reference, but in the case of many species that have low sequence homology with pre-annotated genomes these existing methods do not suffice. We recently developed a method that uses multiple existing homology and machine learning functional annotation methods, based on a community wide assessment of protein-function prediction algorithms [1], for the annotation of some of our micro-algal data.

In this project, aims are to improve consensus-based protein function prediction by:

a) Creating algae specific sub-modules for the existing tools

b) Addition of more function prediction methods to the consensus-based prediction

c) Creating a statistical-based method for combining each methods result into a consensus prediction

a) None of the prediction methods we used considered microalgae when being built. The unique nature of microalgae and their lack of existing annotations mean that the prediction results are sub-par compared to other organisms. Specifically the machine learning method we incorporate needs to be tailored to make it more useful for the annotation of microalgae, which can be done by creating more suitable training sets and adjustment of the biological interpretations.

b) We think that adding more prediction methods will improve the accuracy of our annotation. Currently we incorporate three homology and one machine learning-based method. Addition of protein-protein interaction, gene expression or structure-based methods could prove to be significant additions. To incorporate these we will need to make them available locally on our servers, make them compatible with our existing consensus prediction and test the results.

c) We currently have a working setup for consensus predictions using four different prediction methods. However, we observe a difference in prediction power for each method and as such would like to create a basic statistical framework for the consensus prediction, taking into account the strengths and weaknesses of each method.

Requirements

- Eagerness to learn at least one of: Python, Java, Perl, R
- Willing to learn how to work in a server environment
- Interest in genomics / transcriptomics / proteomics
- Willing to learn a basic level of statisticsContact

Supervisors

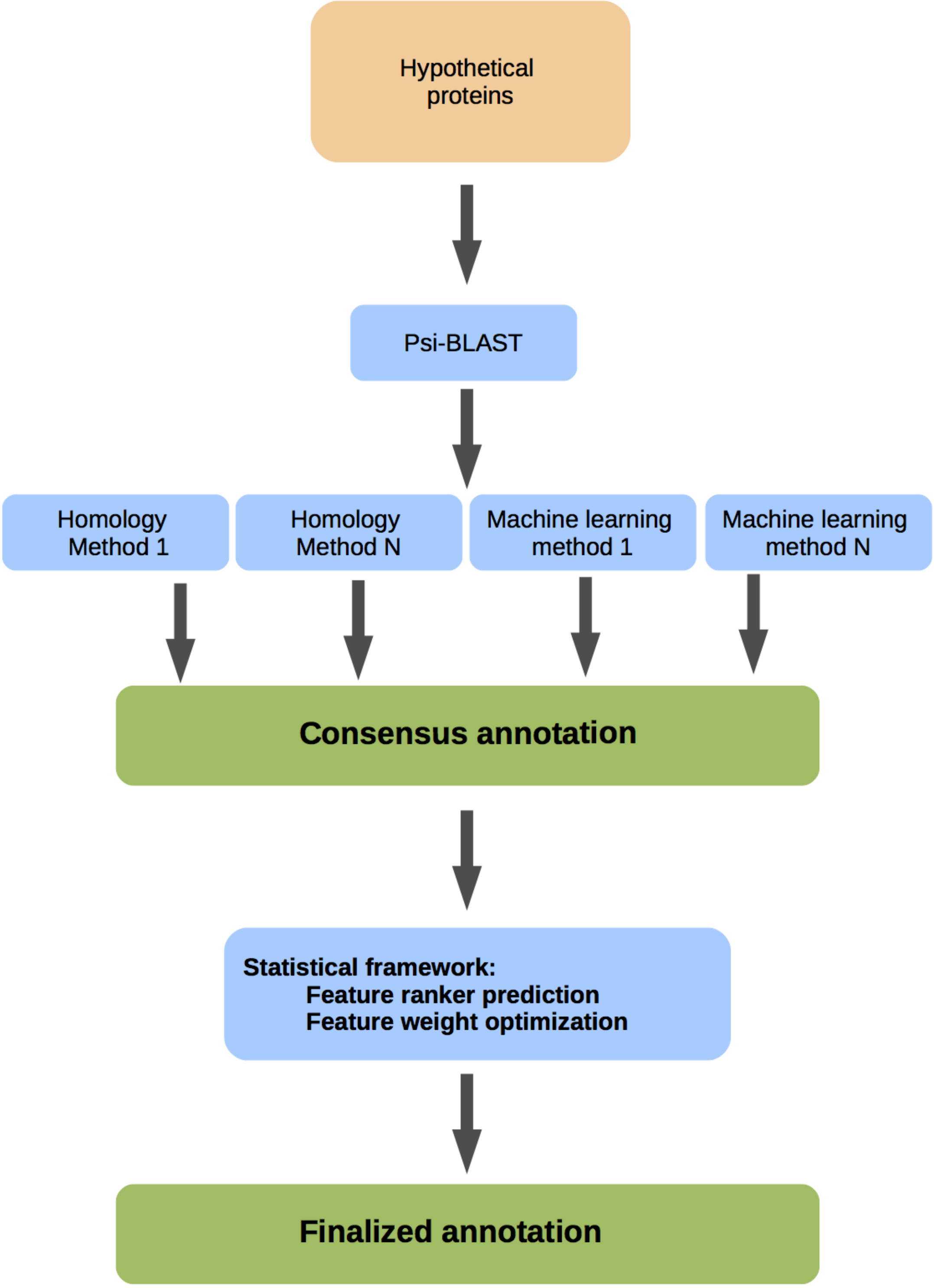
Maarten Reijnders and Peter Schaap

Laboratory of Systems and Synthetic Biology
Wageningen University and Research Center
Dreijenplein 10, Wageningen
6703 HB, the Netherlands

Maarten Reijnders
E: maarten.reijnders@wur.nl
Building 316

Thesis project:

Consensus-based protein function prediction


Hypothetical proteins
Psi-BLAST
Homology Method 1
Homology Method N
Machine learning method 1
Machine learning method N
Consensus annotation
Statistical framework:
Feature ranker prediction
Feature weight optimization
Finalized annotation