

Prediction of RNA and DNA binding sites: preliminary presentation

Quirin Heiss^{1,2} Pandu Raharja ^{1,2} Julian Schmidt ^{1,2}

¹Technische Universität München

²Ludwig-Maximilians-Universität München

November 10, 2016

Outline

- ▶ Binding prediction and CAFA.
- ▶ Background: RNA and RNA binding proteins.
- ▶ Background: datasets.
- ▶ Steps

Definition

- ▶ **RNA/DNA Binding Protein prediction:** given a protein, determine whether a protein is RNA/DNA binding.
- ▶ **RNA/DNA binding site prediction:** given a protein sequence, determine side chains that bind with a DNA/RNA.

CAFA

1. Determine whether protein is RNA or DNA binding.
2. Determine binding site.

Methods: abstraction level

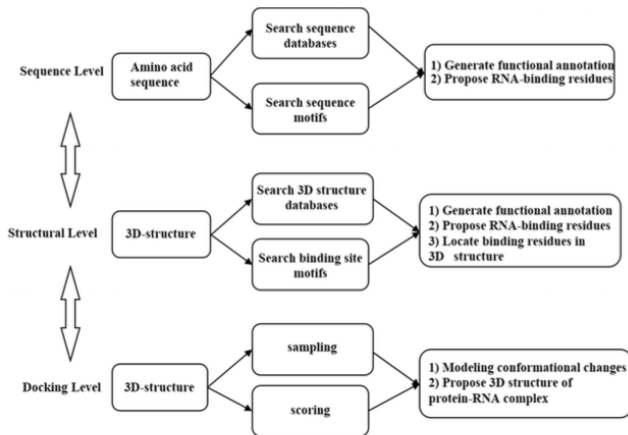


Figure 1. Strategies for RNA-binding site and RBP prediction.

Possible Features

Sequence-based features:

- ▶ **Amino acid composition.**
- ▶ **Sequence similarity**, such as MSA.
- ▶ **Evolutionary invormation**, such as PSSM.
- ▶ **Evolutionary invormation**, such as PSSM.

Structure-based features:

- ▶ **Secondary structure:** experimental (assigned using DSSPcont) or predicted.
- ▶ **Accessible surface area**, in percent (%).

Chemical and physical features:

- ▶ **Hydrophobicity.**
- ▶ **Electrostatic patches.**
- ▶ **Cleft Size.**

Methods: previously used algorithm

- ▶ **Naive Bayes (NB)** classifier.
- ▶ **Support Vector Machine (SVM)**.
- ▶ **Random Forest**.
- ▶ **Neural Network (NN)**.

Statistics

Name	Num
SwissProt (HUMAN)	20120
GO:0003676 (HUMAN)	1248
SwissProt (filtered out GO:0003676)	20005

Statistics (cont.)

Results of redundancy reduction on training set:

Before	After
706	567

Steps

- ▶ Data pre-processing.
- ▶ Models development.
- ▶ Training.
- ▶ Validation.