

Prediction of RNA and DNA binding sites: weekly report

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

¹Technische Universität München

²Ludwig-Maximilians-Universität München

December 15, 2016

Outline

- ▶ Who's presenting what:
 - ▶ Julian: XNA binding prediction
 - ▶ Pandu: XNA binding site prediction
- ▶ Problems.

As always:

- ▶ Scripts and results could be found in:
<https://github.com/raharjaliu/BIFers>

Part I: XNA binding prediction

Troubleshooting

- ▶ no PSI-BLAST hits (matrix all 0)
 - ▶ Warning: sequence Q96IR3 has an empty profile
 - ▶ occurs in all splits \Rightarrow not the problem
- ▶ diag 0 at xxx error
 - ▶ all 606 proteins:
 - diag 0 at 604
 - diag 0 at 605
 - ▶ first 100 proteins:
 - diag 0 at 98
 - diag 0 at 99
 - ▶ only negative proteins (with mock classes): no errors

\Rightarrow some problem with positive proteins (profiles?)

Performance

model	data	k	s	Accuracy	Sensitivity	Specificity	F1
1	2	4	8	0.9422	0.5781	0.9852	0.6789
2	1	4	8	0.9455	0.6094	0.9852	0.7027

Current state

1. Train profkernel: finished (for split.1.dna, split.2.dna, split.0.rna)
2. Predict binding: finished (for existing models)
3. Evaluate predictions for different k and s: finished for

Part II: XNA binding site prediction

First thing first

- ▶ Errors are indeed at P17574 and P17574:
 - ▶ Removed P17574 or P24264: `pp2features.py` finishes.
 - ▶ Included P17574 or P24264: `pp2features.py` crashes.
 - ▶ Running **only** P17574 or P24264: `pp2features.py` crashes.
- ▶ Extracted features could be found in
`/mnt/project/pp2_1516/xrna_raharjaschmidt/
machine_learning/{dna_big.arff,rna_big.arff}`.

Machine Learning Sorcery

- ▶ `pp2features.py` exports to csv: py continuation!
 - ▶ jupyter: interactive analysis and reporting
 - ▶ scikit: features selection + basic ML
 - ▶ pandas: data manipulation
 - ▶ etc
- ▶ Analysis runs natively on Python now → more coming soon.

References