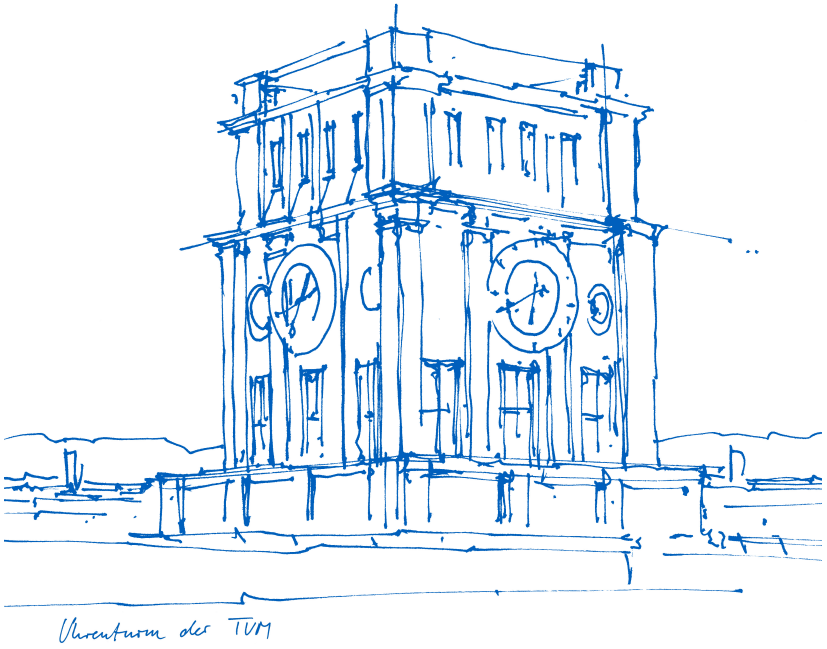
Prediction of Transmembrane Proteins (20 Amino Acid Length)

Pandu Raharja

Nick Lehner

Quirin Heiss

Jun 16, 2016



Background

- Transmembrane proteins essential class of protein:
  - Majority of drugs target are transmembrane protein
  - Account for ~25% of all known proteins
- Identification of structure hard due to biophysical nature of proteins
- A predictive method to identify secondary structure of transmembrane protein would be helpful in research pipeline

Goals

- Development of predicion method for protein transmembrane region:
  - Pipeline development: extract transform load (ETL), training, testing
- Along the way: learning basics of machine learning

Approach

Input:

- Amino acid sequences with annotated secondary structures
- PSI-Blast of the sequences

Methods:

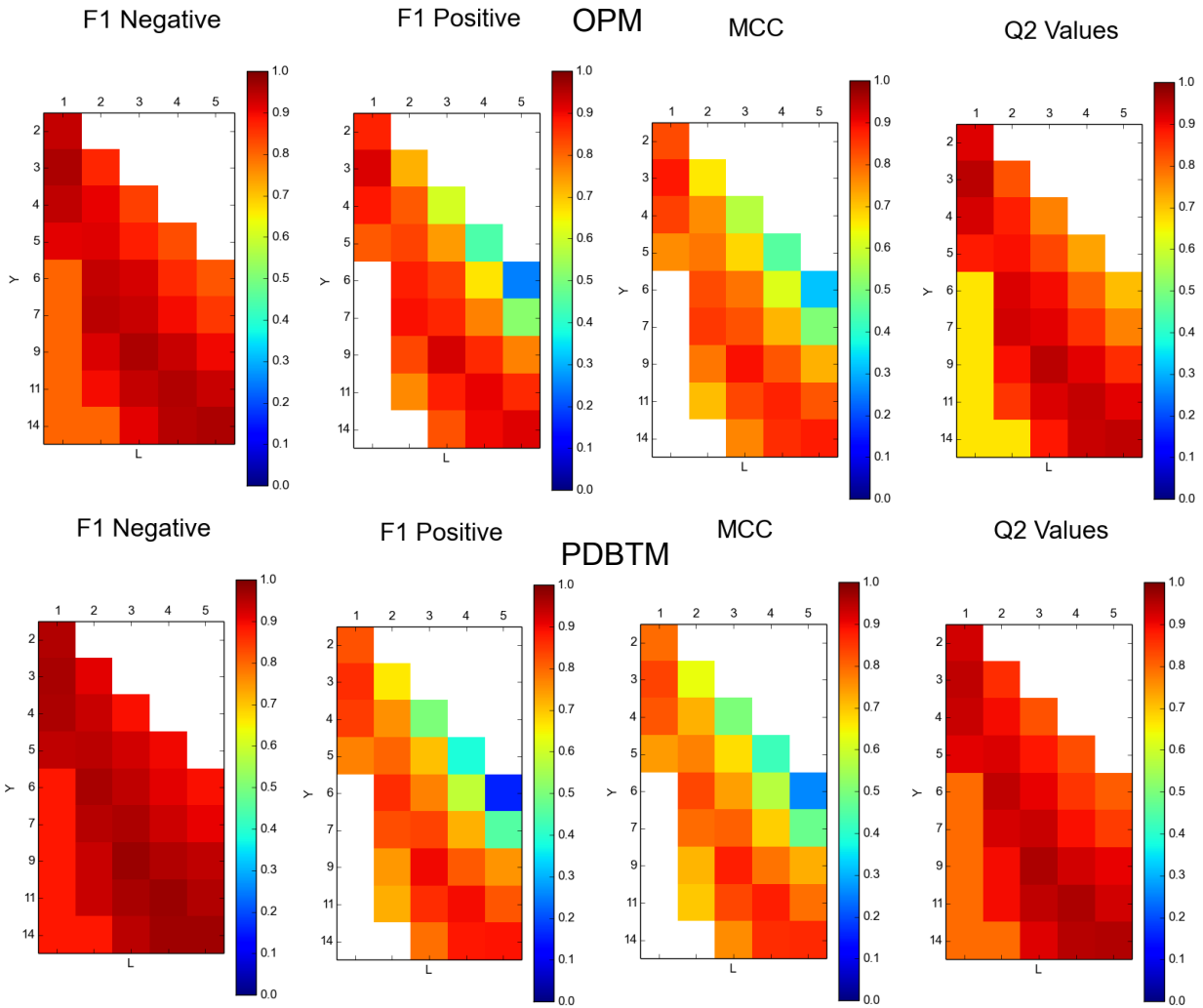
1. Extract fragments:
   - 20 consecutive helical amino acids (positive training set)
   - 20 consecutive non-helical amino acids (negative training set)
2. Extract corresponding PSI-Blast profiles
3. Machine learning pipeline:
   - Application of profile kernel function
   - Cross-validate development set and measure performance
   - Measure performance on independent test set

Data Sets

- Two data sets used: PDBTM & OPM

	Data sets size (positive/negative instances)	
	Development Set	Independent Test Set
TM proteins: 20aa regions	PDBTM: 195/756 OPM: 347/698	PDBTM: 22/85 OPM: 39/78

Results (Development Set)

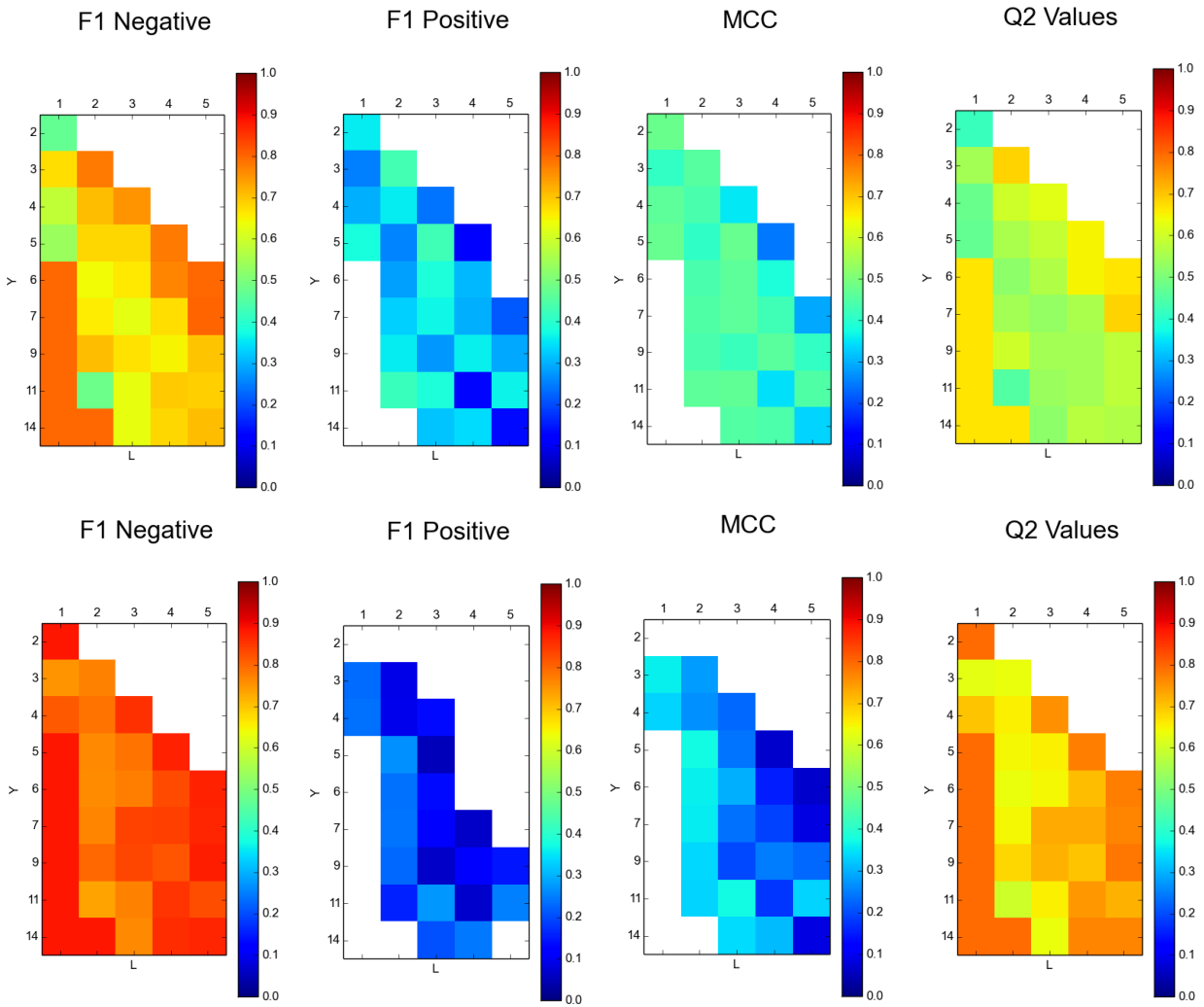


Optimal parameters:

Y=8 L=3
Best Q2, F1 positive, MCC in both OPM and PDBTM data

Y=11 L=4
Best F1 negative in both OPM and PDBTM data

Results (Independent Test Set)



Best performance:

Y=11,L=2
for Q2, F1 Positive, F1 Negative, MCC for PDBTM data

Y=2, L=1
For Q2,F1 Negative in PDBTM and MCC in OPM data

Y=7, L=5
For Q2, F1 Negative in OPM data

Questions?