ALMA MATER STUDIORUM A.D. 1088

UNIVERSITA DI BOLOGNA

PREDICTION OF PROTEIN-PROTEIN INTERACTION SITES WITH A NEW PROBABILISTIC METHOD

Presentata da: Saeideh NAZERI

Relatore:
Chiar.mo Prof. Pier Luigi Martelli

Correlatori:
Dr. Castrense Savojardo

Protein Interactions

- **How proteins come together to construct metabolic and signaling pathways**
- **Determining protein interfaces by experimental methods are expensive and time consuming**
- **particularly problematic for transient complexes**

Computational Methods

protein interaction prediction

if two protein interact?

protein–protein docking

pairwise residue contacts between the two binding protein

protein interface prediction

identify a subset of residues on a protein, which might interact with the presumed binding partner

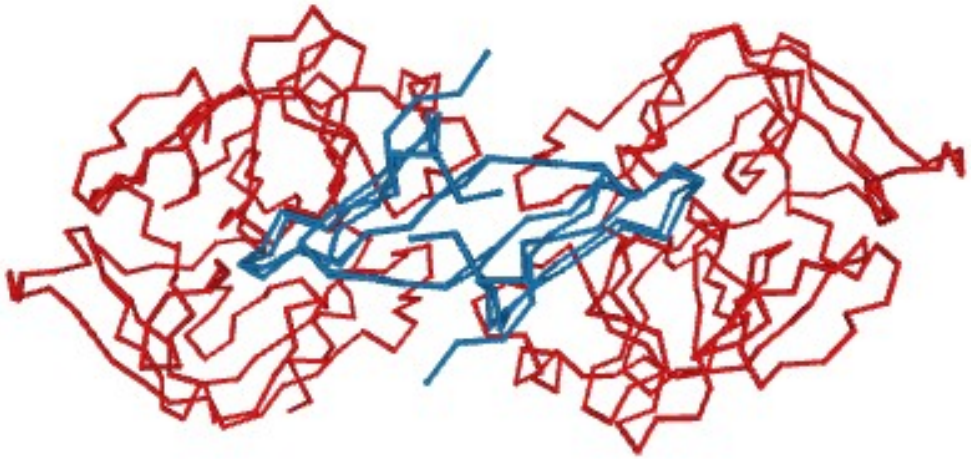
Sequence Labeling Problem

1D6R_I

K P C C D Q C A C T K S N P P Q C R C S D M R L N S C H S A C K S C I C A

S S S B S S S S S S S S S S S S B S B S B S S S S S B B S S S S S S S S S

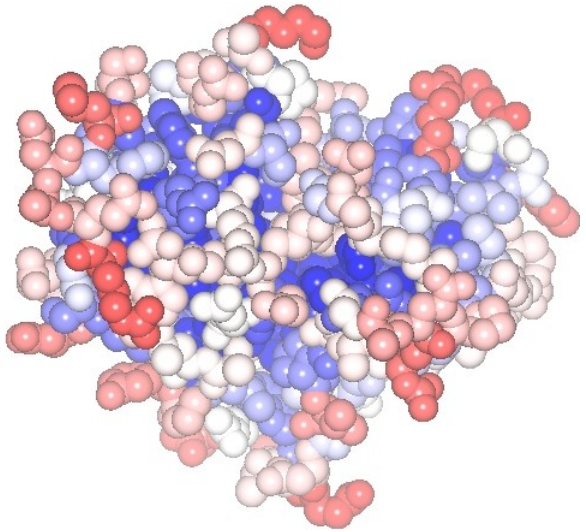
N N N N I I I I I I I I I I I I N I N I N N N N N N N N N N N N N N N N N



Basic Definitions

Surface of Protein

- Exposed residues as those with a Relative Solvent Accessibility (RSA*) above 16% (10% , 5%).
- The higher the threshold, the lower the number of surface exposed residues.

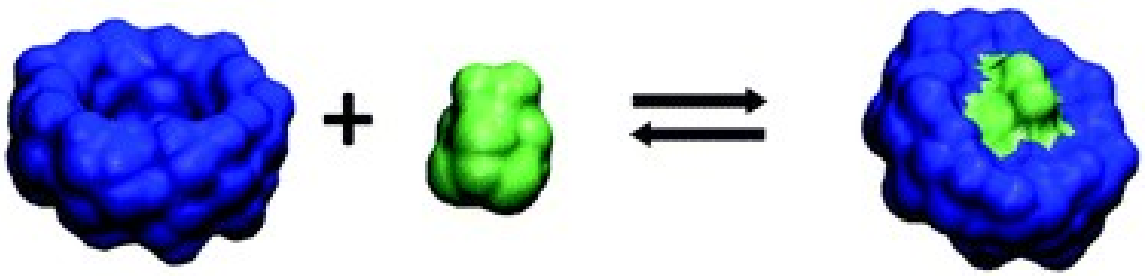


* *Lee, B; Richards, FM. (1971). "The interpretation of protein structures: estimation of static accessibility*"

Basic Definitions

- **Interacting Residue**
  - Differences in the solvent accessible surface area (RSA) when the monomers are separated(using DSSP* application)

  - (ΔRSA) in bound and unbound state for each residue



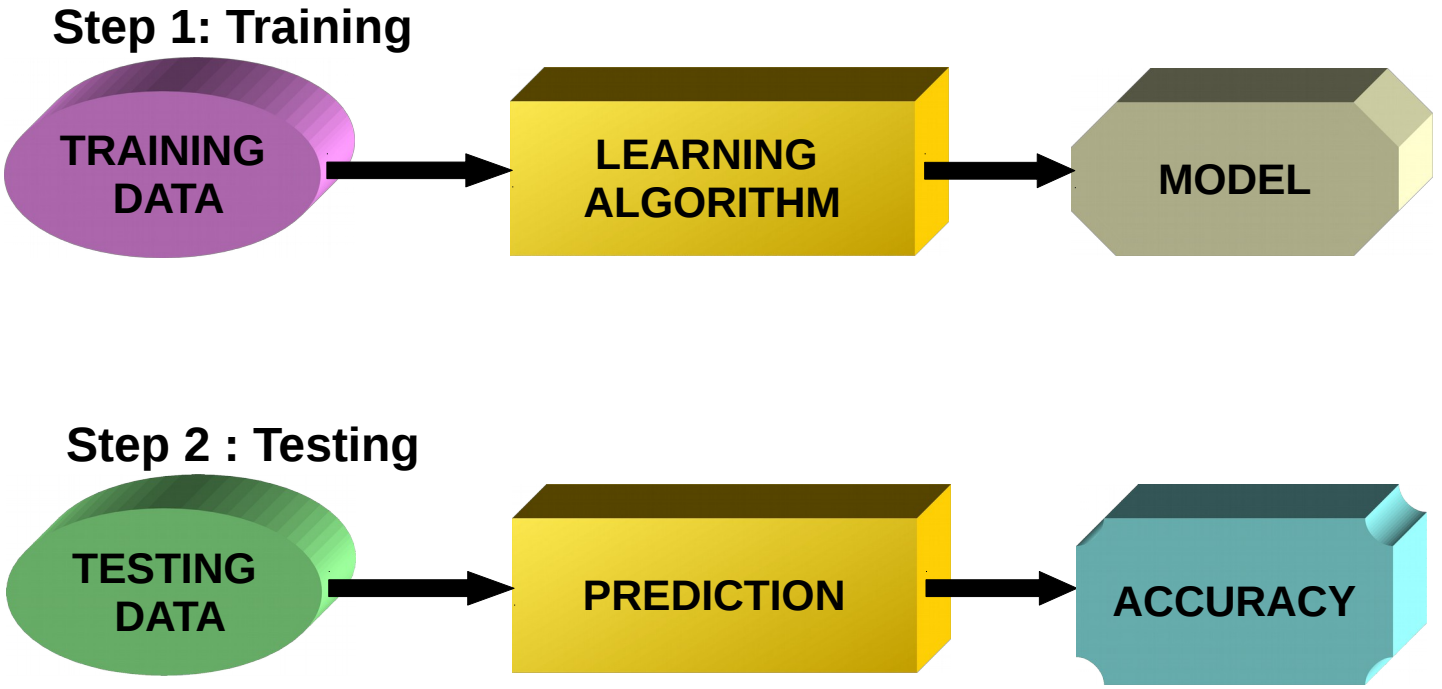
Prediction of Secondary Structural Elements in the Phosphatidycholine-Transfer Protein from Bovine Liver.R. Akeroyd, J. Lenstra, J. Westerman, G. Vriend, K. Wirtz, L. v. Deenen,

Sequence Labeling

Input data(X): Vector of features

each entity often has a rich set of features that can aid classification

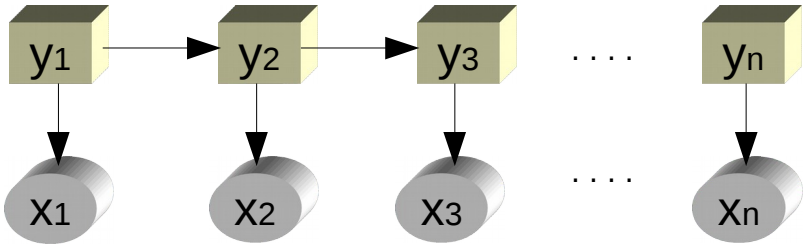
Label(Y): "N" or "I" : Data Classification



Graphical Model

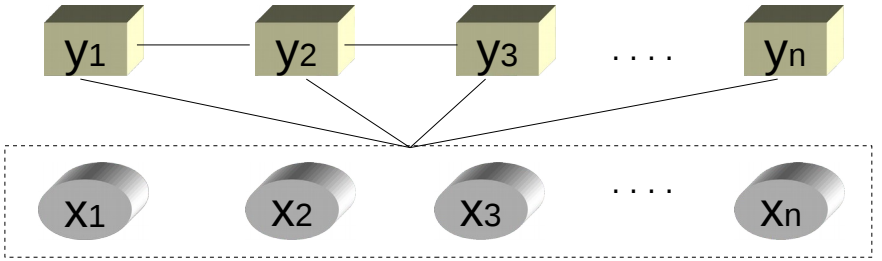
- **Generative**

  Hidden Markov Model(HMM)

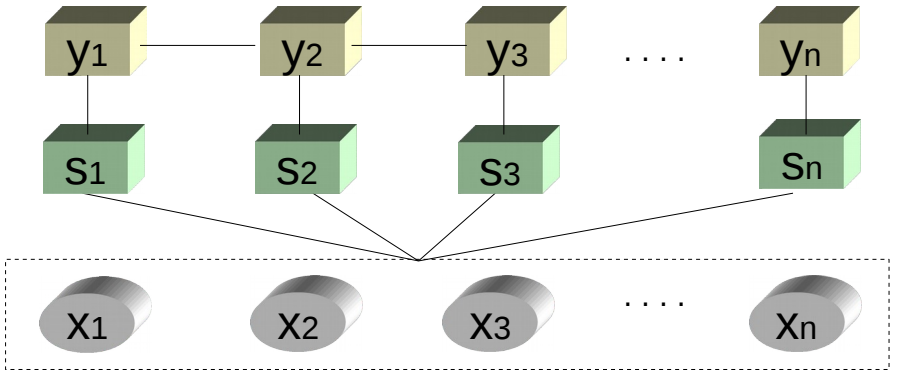


- **Discriminative**

  Conditional Random field(CRF)



Grammatical-Restrained Hidden CRF*

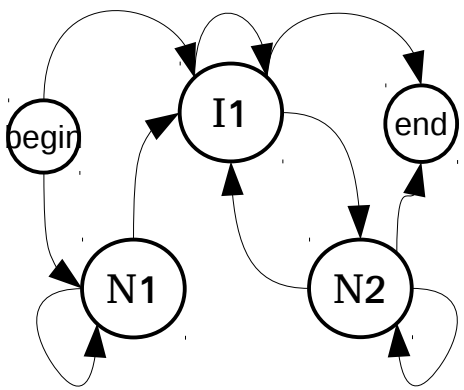


- Include hidden variable into the model
- Provide meaningful results
- Exert our prior knowledge

**Grammatical-Restrained Hidden Conditional Random Fields* for Bioinformatics applications , Piero Fariselli, Castrense Savojardo, Pier Luigi Martelli, and Rita Casadio1 ,2009*

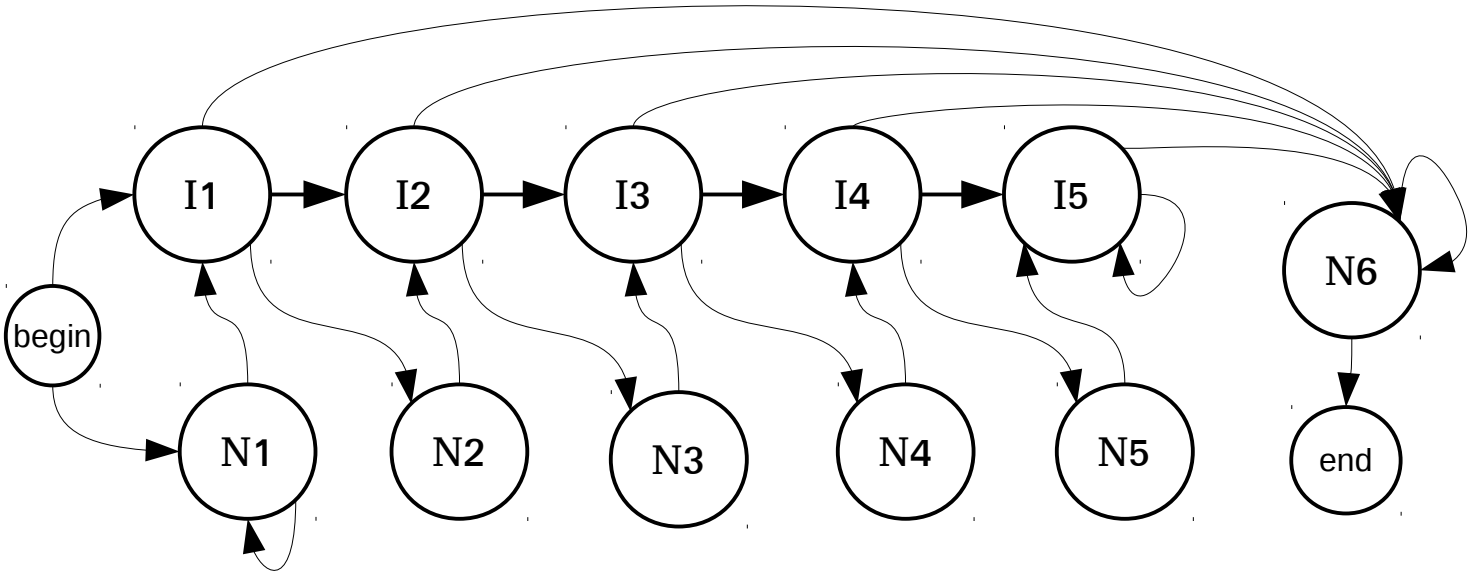
Grammar Definition

Grammar A



Basic Grammar

Grammar B



Grammar for predicting

longer interacting sites

Dataset

Benchmark 5.0

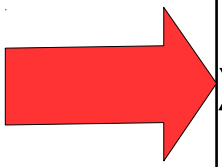
225 protein complexes

High range of interaction types

Bound and unbound state

Non redundant

Experimentally-measured



BM609

Sequence identity <= 25%

X-ray resolution 3.25 Å

At least 50 residues

609 non redundant chain

Features

General Features

- Hydrophobicity
- Multiple Sequence Alignment
- B-Factors
- Electrostatic potential
- Depth and protrusion index
- Simple predicted RSA using SABLE*

Our Method features

- Evolutionary information** of spatial neighbors
- Average of solvent exposure of spatial neighbors

*Accurate prediction of solvent accessibility using neural networks–based regression,R. Adamczak, A. Porollo, J. Meller,2004

**PSSM : profile based on a multiple sequence alignment was created using the PSI-BLAST

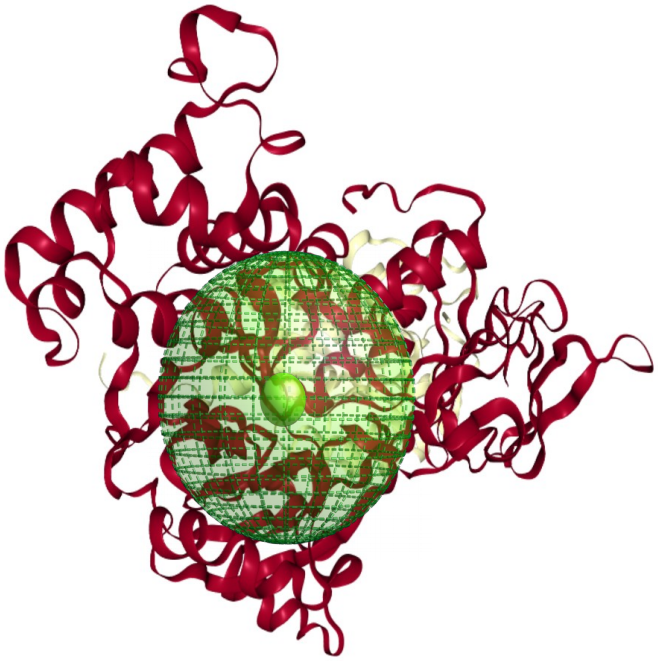
Feature 1: Spatial Neighbors

A sphere with radius of 12Å around each residue and we mark the residues on that area which are on the surface

Neighbors for residue Number 7:
8 , 9 , 10 , 11 , 34 , 254 , 255 , 256 ,282

PSSM for neighbor residues:

8	0	83	2	0	0	13	0	0	0	0	0	1	0	0	0	1	0	0	
9	33	1	1	0	9	1	1	0	0	1	0	4	0	0	0	27	18	0	
10	2	0	2	0	0	0	0	94	0	0	0	0	0	0	0	2	0	0	0	0
11	6	0	0	0	0	1	2	1	0	2	1	0	0	0	1	6	80	0	0	1
34	8	10	3	3	3	5	10	3	3	10	3	5	3	3	8	9	3	7	0	3
254	5	15	26	10	3	4	4	16	3	0	1	5	2	0	0	5	0	0	0	0
256	5	3	3	3	1	4	6	7	1	0	2	8	0	2	33	9	5	1	1	1
282	6	8	3	5	3	3	7	3	3	2	0	35	3	2	6	4	3	0	0	3



Feature 2: RSA difference

Neighbors for residue Number 7:
8 , 9 , 10 , 11 , 34 , 254 , 255 , 256 ,282

Residue	RSA DSSP	RSA SABLE	Difference
8	0.7	0.3	0.4
9	0.0	0.4	-0.4
10	0.3	0.0	3.0
11	0.1	0.0	1.0
34	0.1	0.2	-0.1
254	0.0	0.1	-0.1
255	0.4	0.2	0.2
256	0.1	0.0	0.1
282	0.0	0.0	0.0

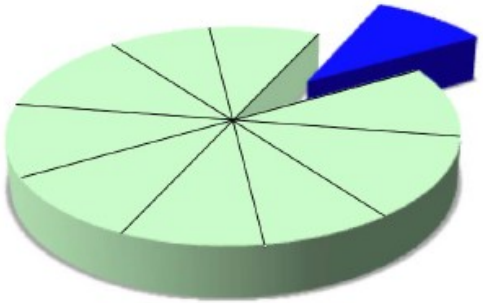
Average = 0.44

Evaluation

10 fold cross validation

repeat 10 times

each part have similar percentage of surface and interacting residues



Scoring Index formulas

True Positive Rate

$$\frac{TP}{TP+FN}$$

Precision

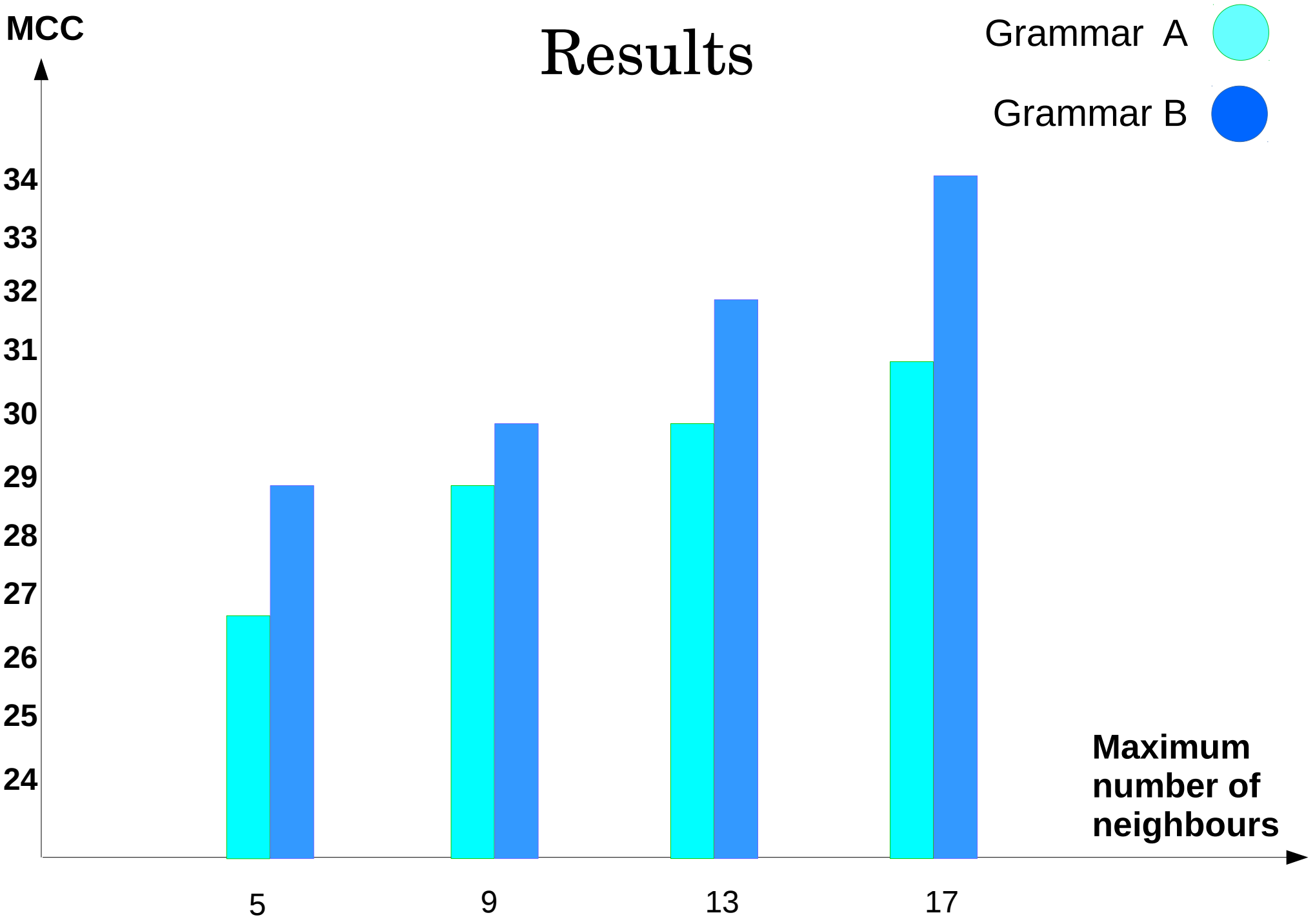
$$\frac{TP}{TP+FP}$$

Accuracy

$$\frac{TP+TN}{TP+FP+TN+FN}$$

MCC

$$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(FP+FN)(TN+FN)}}$$


Results
MCC
Grammar A
Grammar B
34
33
32
31
30
29
28
27
26
25
24
Maximum number of neighbours
5
9
13
17

Results Detail

DATASET	TPR	PPV	ACC	MCC
Window 17	32%±0.01	67%±0.02	76%±0.04	34%±0.02*
Window 13	30%±0.07	66%±0.08	76%±0.01	32%±0.06
Window 9	28%±0.06	66%±0.01	75%±0.06	30%±0.08
Window 5	25%±0.04	65%±0.06	75%±0.03	29%±0.03

Grammar B on BM_609 Dataset

Comparison

Predictors	TPR	PPV	ACC	MCC
Pseudo-meta	69%±0.02	41%±0.07	88%±0.07	48%±0.01
PredUs	70%±0.01	30%±0.02	30%±0.02	38%±0.03
eFindSite PPI	39%±0.06	45%±0.09	90%±0.05	37%±0.05
Best Result	**32%±0.01**	**67%±0.02**	**76%±0.04**	**34%±0.02**
cons-PPISP	27%±0.09	33%±0.08	88%±0.08	24%±0.07
SPPIDER	34%±0.00	20%±0.08	82%±0.07	17%±0.03
ProMate	52%±0.06	21%±0.00	68%±0.04	16%±0.05
WHISCY	13%±0.00	33%±0.04	90%±0.00	16%±0.04
PIER	06%±0.06	34%±0.02	90%±0.06	11%±0.08
VORFFIP	53%±0.01	33%±0.07	57%±0.09	11%±0.07
PSIVER	64%±0.05	11%±0.08	54%±0.06	10%±0.03
InterProSurf	43%±0.05	16%±0.03	67%±0.07	10%±0.00

Conclusion

- Stable performance given group of proteins
- True interface is usually in common by one or more predictors
- There is high capacity of improvement by expanding features and defining more sophisticated grammars

Thanks

for your attention