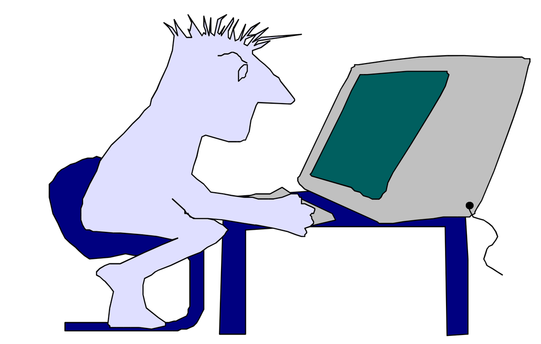
Protein Sequence Classification in Proteome Analyst

Zhiyong Lu*, Xiaomeng Wu and Russ Greiner
University of Alberta
zhiyong@cs.ualberta.ca
http://www.cs.ualberta.ca/~bioinfo/PA



labeled data

Protein Class	Protein Sequence
Cell_envelope	mtvatgdpa…
Purines	mtvatgdpv…
Replication	mrseksltl…
….	….

SwissProt Database

Go Psiblast

```
sp|P00561|AK1H_ECOLI  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ...  1114  0.0
sp|P27725|AK1H_SERMA  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ...  1074  0.0
sp|P49079|AKH1_MAIZE  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ...  1023  0.0
sp|P49080|AKH2_MAIZE  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ...  993  0.0
sp|P44505|AKH_HAEIN  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D...  981  0.0
sp|P37142|AKH_DAUCA  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D...  980  0.0
sp|P57290|AKH_BUCAI  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D...  974  0.0
sp|P00562|AK2H_ECOLI  BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ...  818  0.0
```

```
> unknown 1
->AK1H-Ecoli
```

similar sequences for unknown-1

unknown sequences that have high similarity to some sequences in the SwissProt database

most similar sequences for unknown-1

```
LOCUS       AK1H_ECOLI      820 aa      linear   BCT 16-OCT-2001
DEFINITION  Bifunctional aspartokinase/homoserine dehydrogenase I (AKI-HDI)
            [Includes: Aspartokinase I ; Homoserine dehydrogenase I ].
ACCESSION   P00561
PID         g113539
VERSION     P00561  GI:113539
DBSOURCE    swissprot: locus AK1H_ECOLI, accession P00561;
            class: standard.
            extra accessions:Q47659,created: Jul 21, 1986.
            sequence updated: Jul 21, 1986.
            annotation updated: Oct 16, 2001.
xrefs (non-sequence databases): EcoGene EG10998, InterPro
            IPR002912, InterPro IPR001048, InterPro IPR001341, InterPro
            IPR001342, Pfam PF00696, Pfam PF01842, Pfam PF00742, PROSITE
            PS00324, PROSITE PS01042
KEYWORDS    Transferase; Kinase; Oxidoreductase; Threonine biosynthesis; NADP;
            Allosteric enzyme; Multifunctional enzyme; Complete proteome.
SOURCE      Escherichia coli.
...
```

Feature extractor

Seq. #	Class	Transferase	Kinase	NADP	Chloroplast	IPR001342	Threonine	IPR001048	Transit
1	C	1	1	1	0	0	0	0	0
2	C	1	1	1	1	0	0	0	0
3	P	0	0	1	0	0	1	0	1
4	P	0	0	1	0	0	0	1	1
5	R	1	1	1	0	0	1	0	0
6	R	0	1	1	0	0	1	0	0
7	M	0	0	1	0	1	1	0	0
8	M	0	1	1	1	0	0	0	0

Tokenizer Start

Tokenizer Output

thousands of features

Learning procedure

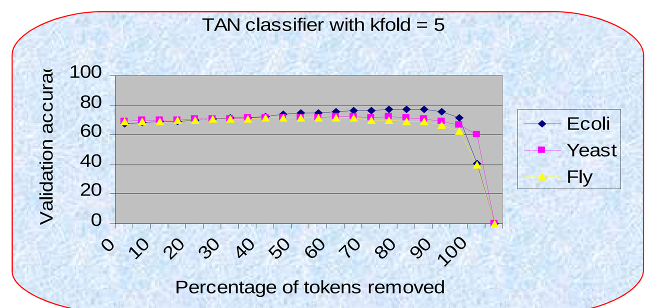
Tokenizer → **Feature Ordering** → [**Feature Selection: Wrapper model** ⇄ **Build Classifier**] → **Classifier**

Information Content:

$$W_j = \sum_v \sum_c P(y=c, F_j=v) \log \frac{P(y=c, F_j=v)}{P(y=c)P(F_j=v)}$$

Feature	Information content
Transit	0.406
Transferase	0.352
Kinase	0.352
Chloroplast	0.156
IPR001048	0.147
IPR001352	0.102
Threonine	0.102
NADP	0

TAN+Wrapper



Machine Learning !

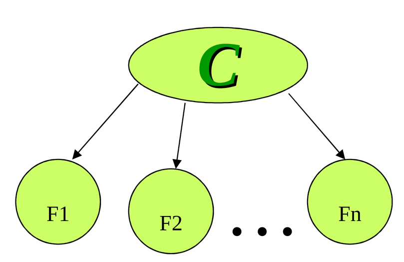
unlabeled data

Protein Class	Protein Sequence
?	mtvgtpdav…
?	mtvdpgtat….
?	mtvtagpltl…
…	…

Tokenizer → **Classification** → **(predicted) Protein Class labels**

Naïve Bayesian Net (NB)

- **Assumption:**
  features are conditionally independent, given class labels
- **Structure:**
  1 level tree
  class labels — root
  features — leaf nodes



Input: feature vector F (F1,F2,…,Fn)

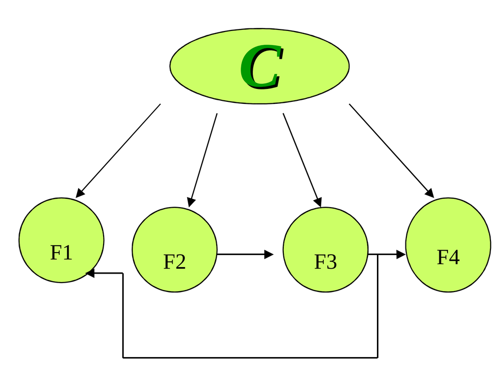
$$P(F|C) = \prod_{i=1}^{n} P(Fi|C)$$

Prediction:

$$class(F) = \arg\max_h P(C)P(F|C) = P(C)\prod_{i=1}^{n} P(Fi|C)$$

Tree Augmented Naïve Bayesian Net (TAN)

- **Assumption:**
  allow some additional edges between features for simple correlation between the features
- **Structure:**
  approximate the interactions among features using a tree structure among features, as well as link from class to each feature



Input: feature vector F = (F1, F2, … , Fn)

Conditional Mutual Information between every two features F1 and F2, given C:

$$\sum_{f1,f2,c} P(f_1, f_2, c) \log \frac{P(f_1, f_2|c)}{P(f_1|c)P(f_2|c)}$$

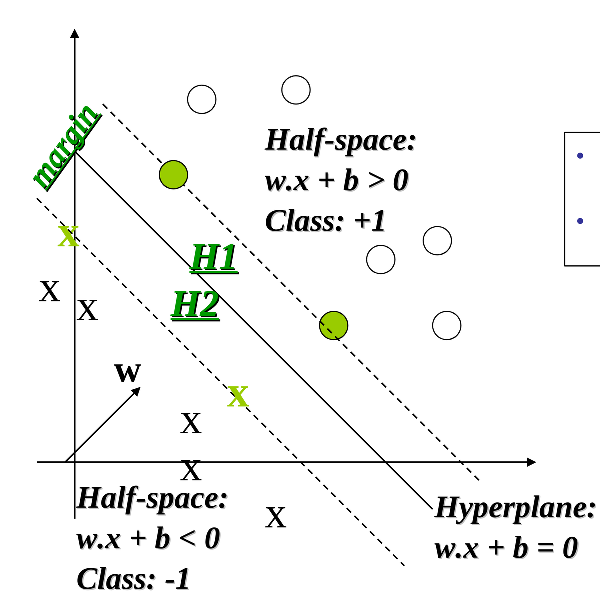
Algorithm for learning structure (links between features) : Chow and Liu, 1968

Prediction:

$$class(F) = \arg\max_c \{P(c)\prod_{i=1}^{n} P(f_i \mid p(f_i), c)\}$$

Support Vector Machine (SVM)

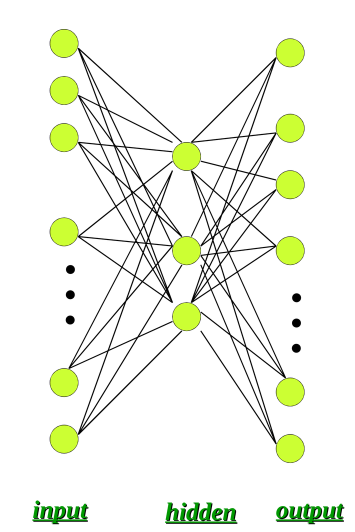
- Input vectors are separated into positive vs. negative instance
- Data points that lie on the margin are "support vectors"
- Map to new feature space such as polynomial function and RBF



- Perceptron: Linear separation of the input space
- h(x) = sign(<w, x> + b)

Artificial Neural Net (ANN)

- Practical for learning real-valued and vector-valued functions over continuous and discrete-valued features
- Robust to noise in training data
- Successful application in many other fields



Input node:
feature vector F (F1,F2,…,Fn)

Hidden node:
one layer, fully connected
Backpropagation algorithm

Prediction:
each output node for one class

Logistic Regression (LR)

- Aims to produce smallest empirical classification error
- Gradient-descent algorithm is used to set parameters
- Learning algorithm descends in the direction of total derivative, given a set of training data

Classification Error:

$$Err(h) = P_{<e,c>}(h(e) \neq c)$$

Empirical Classification Error:

$$Err'(h) = \frac{1}{|S|} \sum_{<e,c> \in S} (h(e) \neq c)$$

Logistic Regression = Discriminative Learning of NB

Learn the CPTable entries for the given NB structure to produce larger empirical LCL score, hence, smaller error

Log conditional likelihood(LCL):

$$LCL(h) = \sum_{<e,c>} P(e;c) \log P_h(c|e)$$

Empirical LCL:

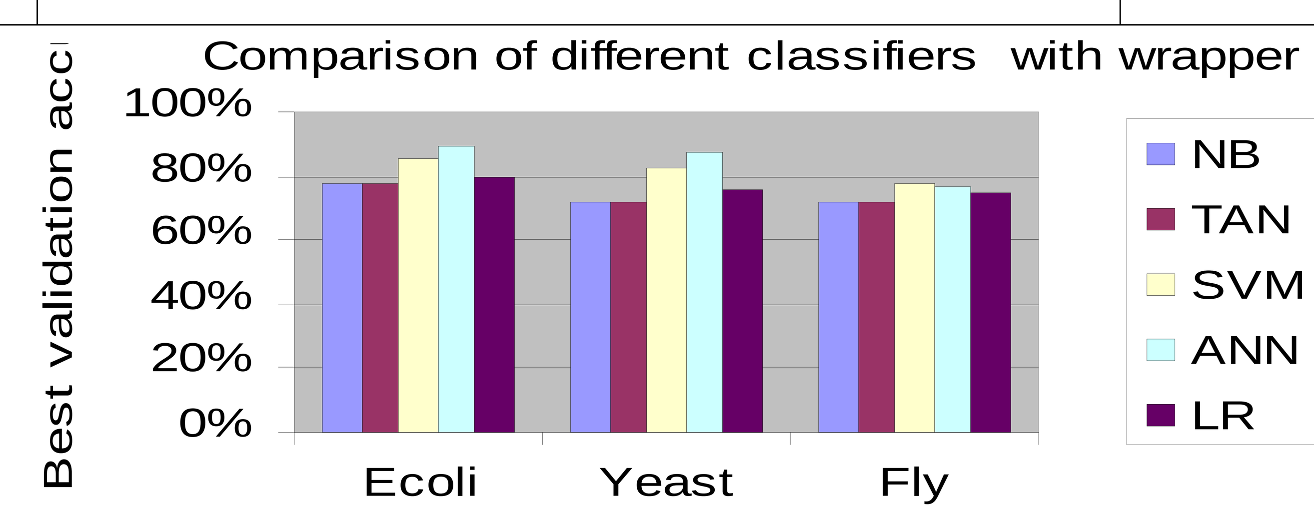
$$LCL'(h) = \frac{1}{|S|} \sum_{<e,c> \in S} \log P_h(c|e)$$

Initial CPTable
For each training data, calculate partial derivative
Sum up to get a total derivative
Gradient-descent algorithm to update each CPTable entry
Get better conditional likelihood

"MORE ACCURATE" !

Experiments:
- Three data sets : Ecoli, Yeast, Fly
- Evaluate each classifier using 5-fold cross validation

Results:
- Feature selection (wrapper model) improves accuracy
- ANN and SVM give best performance



Acknowledgement:
Dept. of Computing Science
PENCE
Dr. Duane Szafron, Dr. Paul Lu
James Redford, Roman Eisner

This work was partially funded by grants from PENCE and NSERC