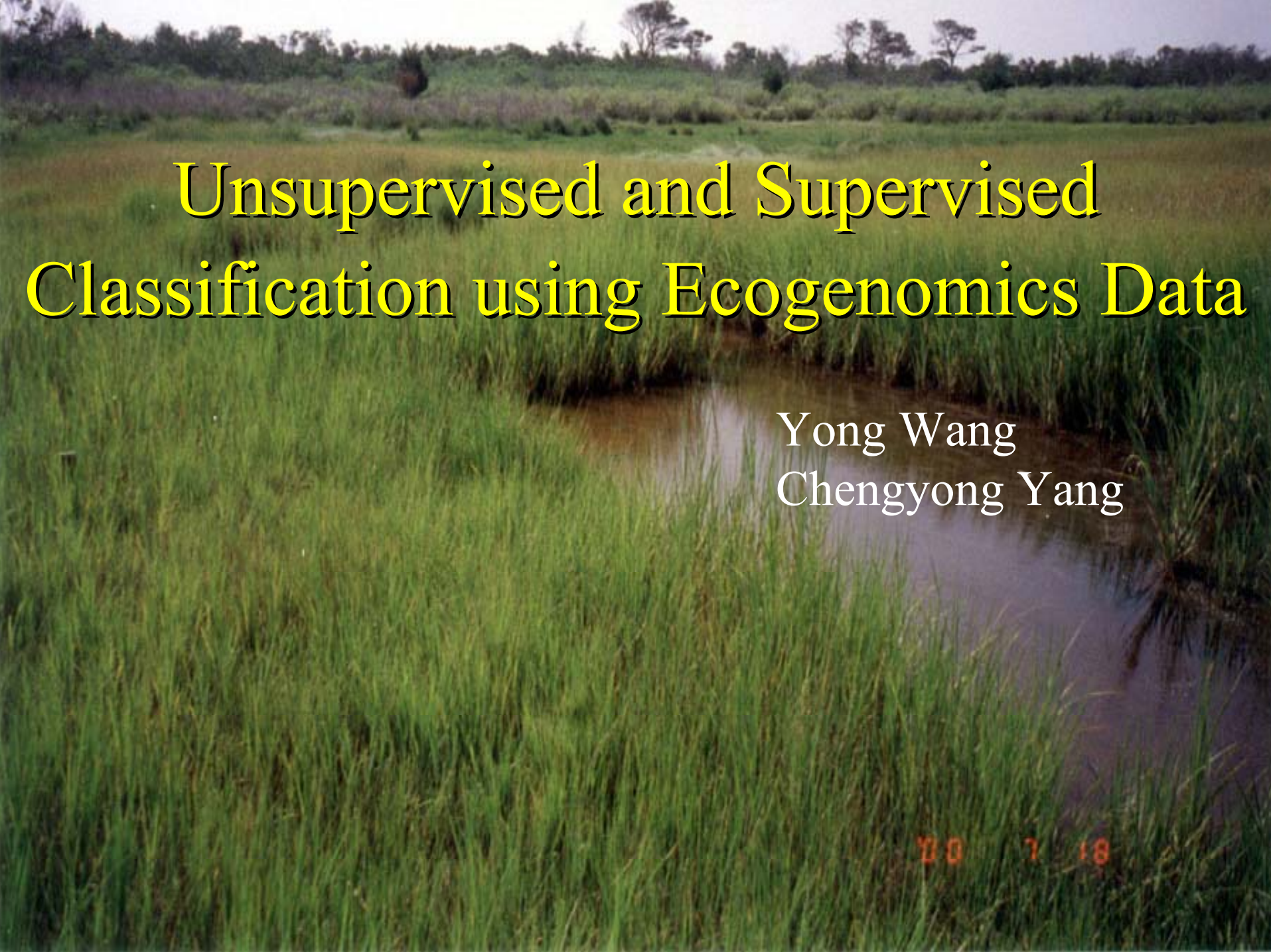


Unsupervised and Supervised Classification using Ecogenomics Data

Yong Wang
Chengyong Yang

Outline

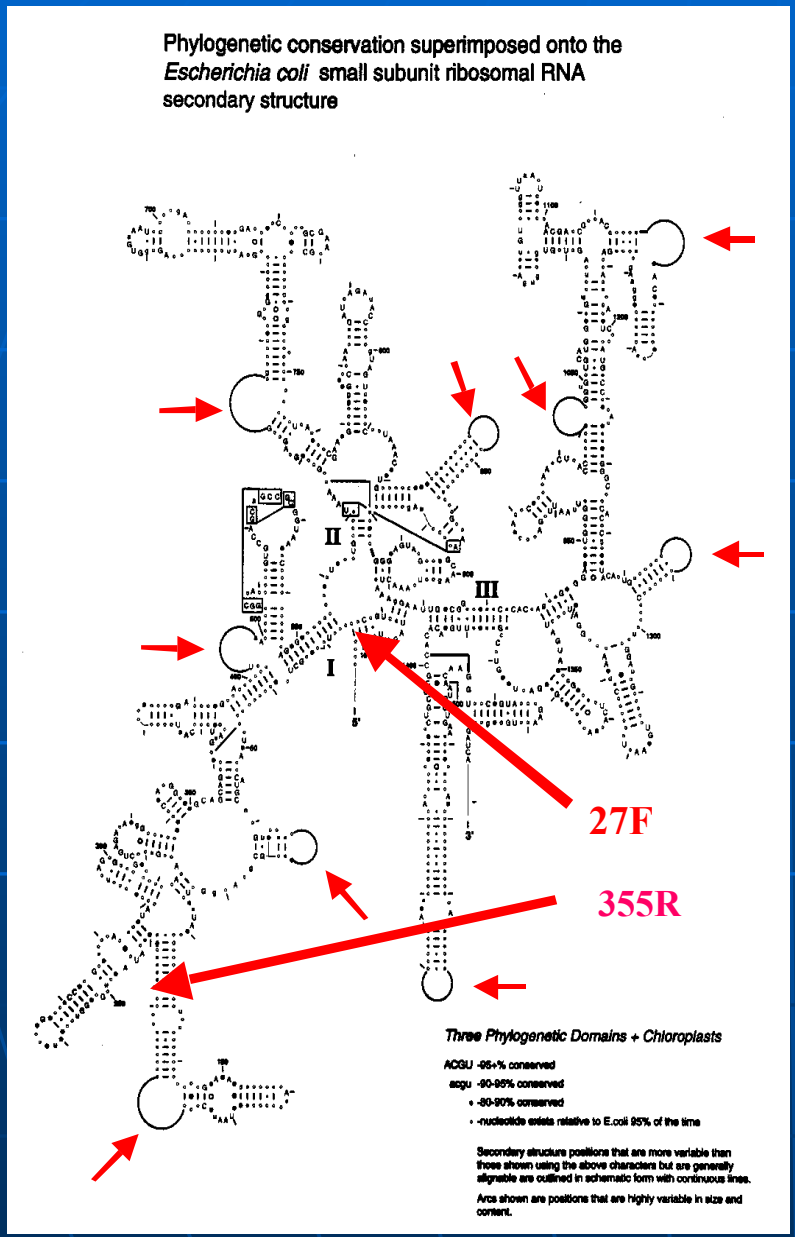
- Objective
- Introduction
- Unsupervised classification
- Supervised classification
- Discussions
- Acknowledgements

Objective

To characterize and classify the microbial communities from different environments using 16S rRNA

Eco-Informatics

- Definition: a broad interdisciplinary science that incorporates both conceptual and practical tools for the understanding, generation, processing, and propagation of ecological information
- Five areas that computer scientists can contribute to Eco-informatics:
  1. Data and collections management.
  2. Wireless communications.
  3. Hi-performance computing, modeling, simulations.
  4. Scientific visualizations.
  5. Data analysis, mining, decision support.

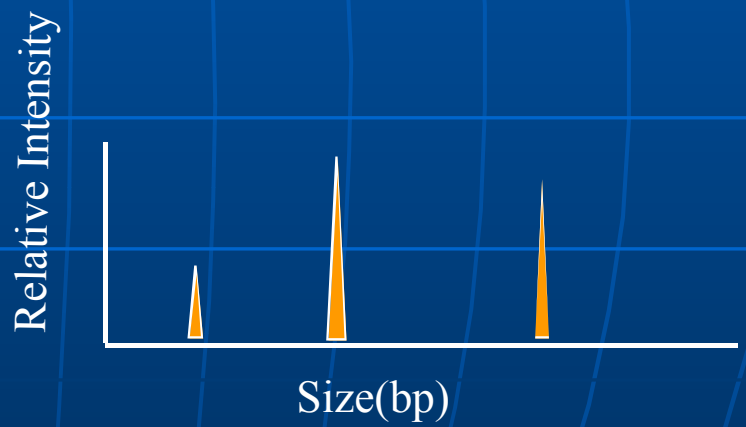


Amplicon Length Heterogeneity Fingerprinting (ALH)

OTU 1 *

OTU 2 *

OTU 3 *



Peak area ~ Abundance

ALH Profile Data

- Data: abundance of various length fragments
- Problems: high dimensional, noisy, very small or very large datasets

Clustering Tools

- Unsupervised: learning without a teacher
  1. Hierarchical clustering
  2. K-Means clustering
  3. Self-Organizing Maps
- Supervised: learning with a teacher

  Suppport Vector Machines

Review of Hierarchical Clustering

- Hierarchical Clustering
  - Greedy, neighbor-joining clustering
  - Two hierarchical methods used:
    1. squared Euclidean distance
    2. correlation coefficient
  - Produce dendrograms: graphic summery of data, rather than description of results
  - Small change in data or hierarchical methods may result in great change to outputs
  - If misled at early stage, prone to errors

Review of K-Means and SOM

- K-Means Clustering
  - Need to specify the number of clusters(i.e. K).
  - Can recover from early mistakes
  - K changes, the cluster membership can change in arbitrary way.
- Self-Organing Maps (SOM)
  - Similar to K-Means
  - Provide information about potential “neighborhood”
- More details about these methods: Hastie and Tibshirani, The Elements of Statistical Learning

Idaho Data

- Four types of soils representing four different management practices
  - CT: conservation tillage
  - NSB: natural sage brush
  - MP: moldboard plow
  - IP: irrigated pasture
- Varying depths in the range of 0-5cm, 5-15cm, or 15-30cm
- Three different subtypes for each sample
- Two or three replicate data for each sample
- Primers targeting V1, V2, and V1 & V2
- Totally 29 samples, 69 dimensions.

Results: Hierarchical Clustering

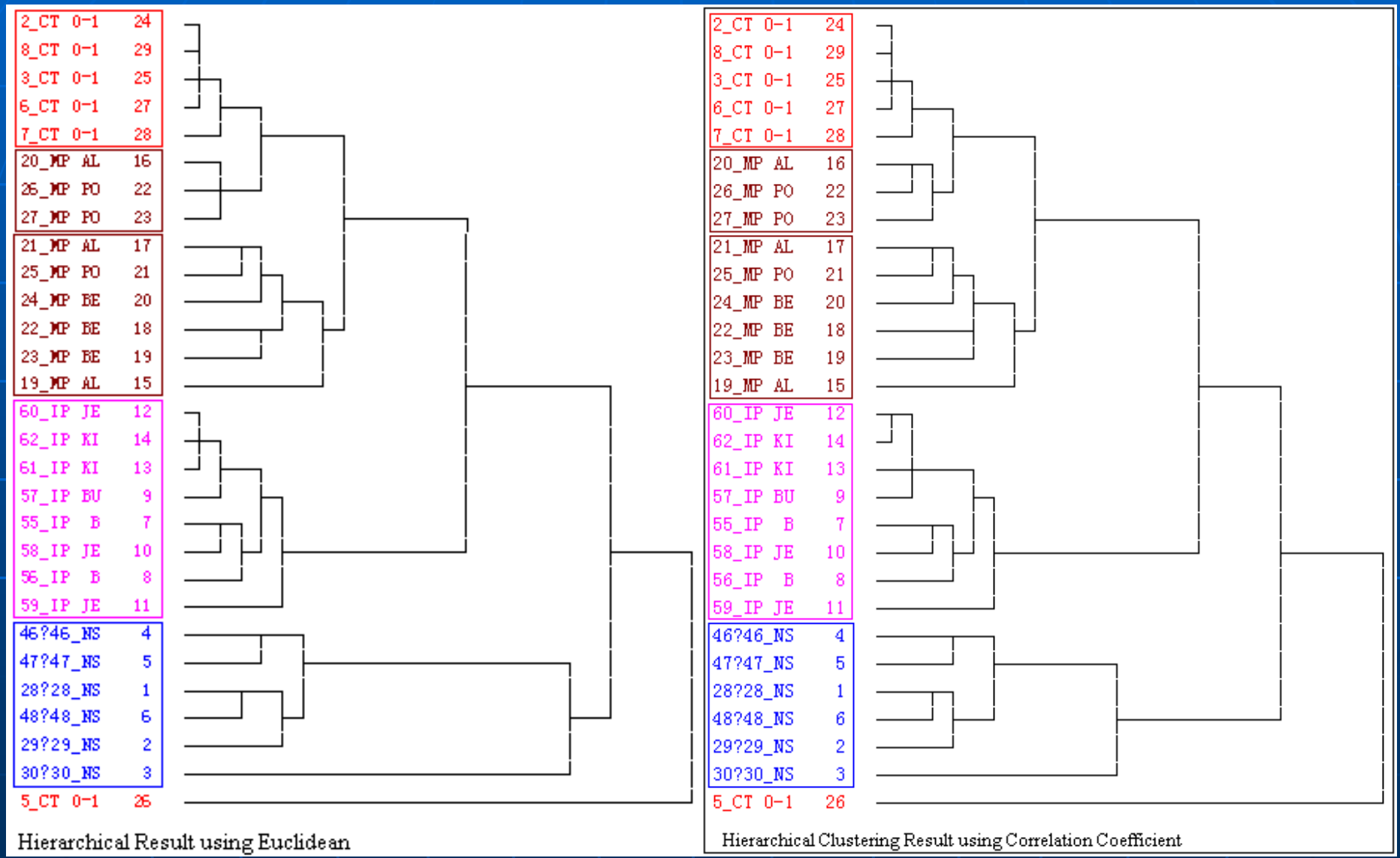


Figure 1: Hierarchical Results

Results: K-Means Clustering

K = 4

Cluster #0	Cluster #1	Cluster #2	Cluster #3
19_MP ALF 1-1	29?29_NSB HW 0-5 1-2	6_CT 0-15 2-3	61_IP KIMBERLY 1-1
23_MP BEAN 1-2	28?28_NSB HW 0-5 1-1	3_CT 0-15 1-3	59_IP JEROME 1-2
21_MP ALF 1-3	48?48_NSB KB 0-5 1-3	2_CT 0-15 1-2	62_IP KIMBERLY 1-2
22_MP BEAN 1-1	47?47_NSB KBR 0-5 1-2	7_CT 0-15 3-1	57_IP BUHL 1-3
25_MP POTATO 1-1	30?30_NSB HW 0-5 1-3	5_CT 0-15 2-2	58_IP JEROME 1-1
24_MP BEAN 1-3	46?46_NSB KB 0-5 1-1	8_CT 0-15 3-2	55_IP BUHL 1-1
26_MP POTATO 1-2			56_IP BUHL 1-2
27_MP POTATO 1-3			60_IP JEROME 1-3
20_MP ALF 1-2			

Figure 2 Four Clusters generated by K-Means Method. There were no outliers in the clusters generated

Results: K-Means Clustering

K = 12

Cluster #	Soil types and their counts
0	MP-ALH(2) MP-Bean(3)
1	-
2	NSB-KBR(1)
3	CT-2(1)
4	NSB-HW(2) NSB-KB(2)
5	NSB-HW(1)
6	-
7	CT-1(1) CT-2(2) CT-3(2)
8	MP-Potato(1)
9	IP-BUHL(1) IP-JEROME(1)
10	IP-BUHL(2) IP-JREROME(2) KIMBERLY (2)
11	MP-Potato(2) MP-ALH(1)

Figure 3: Twelve clusters generated by a K-Means clustering method. Every cluster only contained samples from one soil management type. But it does not differentiate subtypes.

Results: SOM Clustering

1 by 4

NSB(6)	IP(6)	MP(3) CT(6)	IP(8)

Figure 4(a): The contents of the 4 clusters from using a 1 by 4 SOM

Results: SOM Clustering

2 by 2

NSB(6)	MP(3) CT(6)
MP(6)	IP(8)

Figure 4(b): The contents of the 4 clusters from using a 2 by 2 SOM

Results: SOM Clustering

3 by 4

NSB HW(3) NSB KB(3)		MP ALF(2) MP BEAN(1)
IP BUHL(3) IP JEROME(3) IP KIMBERLY(2)		MP ALF(1) MP POTATO(2)
	CT-2(1)	CT-1(2) CT-2(1) CT-3(2)

Figure 4(c): SOM results using 3 by 4

Results Evaluation: Entropy

- Entropy: a measure of the information content or "uncertainty" of a group.

$$\text{Entropy} = \sum_{j=1}^{n} \left(\sum_{i=1}^{k} \left| P_i^j \times \left(\log P_i^j \right) \right| \right)$$

- The smaller the entropy, the better the result
- Outliers make entropy very large, and cannot satisfactorily indicate the actual clustering results

Results Evaluation: Robust Entropy

- Throw some percent of “bad” samples
- Recalculate Entropy
- Improvement Rate = (1 – RE / Entropy) x 100%

Results Evaluation: Sample

Cluster #	Soil types and their counts
0	MP-ALH(2) MP-Bean(3)
1	-
2	NSB-KBR(1)
3	CT-2(1)
4	NSB-HW(2) NSB-KB(2)
5	NSB-HW(1)
6	-
7	CT-1(1) CT-2(2) CT-3(2)
8	MP-Potato(1)
9	IP-BUHL(1) IP-JEROME(1)
10	IP-BUHL(2) IP-JREROME(2) KIMBERLY (2)
11	MP-Potato(2) MP-ALH(1)

Entropy = 6.996

29 samples, eliminate 3, 10%

RE (10%) = 4.556

Improvement Rate = 34.8%

K-Means Clustering Summary

Data	Num of Samples	Dimensions	Cluster Num	Entropy	Entropy RE(0.1)	Improvement Rate(%)
Idaho	29	69	4	0	0	-
			12	4.996	4.556	35
SC Manure Treated	13	81	5	1.899	0.811	57
SC Manure Treated vs. Unmanure Treated	60	49	2	0.297	0	100
			11	14.186	9.265	35
Virginia(by Site)	281	82	7	14.071	10.544	25
			9	16.383	11.589	29
Virginia(by Time)			7	11.907	8.830	26
			9	14.184	10.552	26
Average	-	-	-	-	-	45.2

Figure 5: K-means clustering summary

SOM Clustering Summary

Data	Num of Samples	Dimens--ions	Col by Row	Average Entropy	Average Entropy RE(0.1)	Improvement Rate(%)
Idaho	29	69	1 x 4	0.918	0	100
			2 x 2	1.441	0.459	68
			3 x 4	6.295	4.068	35
SC Manure Treated	13	81	1 x 5	1.500	0.918	38
			2 x 3	1.500	0.918	38
SC Manure Treated vs. Unmanure Treated	60	49	1 x 2	0.149	0	100
			3 x 4	11.805	8.097	31
Virginia(by Site)	281	82	3 x 3	16.540	12.361	25
Virginia(by Time)			3 x 3	15.036	10.281	32
Average	-	-	-	-	-	51.9

Figure 6: SOM clustering summery

Conclusion of Unsupervised Clustering

- Most of data sets are clustered well according to management practices
- Exception Virginia Data
- Evaluation is difficult for the data without a priori knowledge.

Short Review for SVMs

- The SVM problems: finding one hyperplane
- For linearly-inseparable problems: feature space
- Maximal margin classifier: minimizing the risk of overfitting

Kernel Functions

- Linear: $K(X,Y) = (X \bullet Y + 1)^d$ d=1
- RBF: $K(X,Y) = \exp(-\|X-Y\|^2/2\alpha^2)$
- Sigmoid: $K(X,Y) = \tanh(\omega(X \bullet Y)+\theta)$

Test Methods

- Jackknife: Singling out one for test and training the remaining
- Jackrep: Singling out all replicates for test and training the remaining

$$\text{total accuracy} = \frac{\sum_{i=1}^{k} p^{(i)}}{N},$$

$$\text{accuracy } (i) = \frac{p(i)}{\text{obs}(i)},$$

$$\text{MCC}(i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}}$$

Idaho – top soil

Location	Linear		RBF		Sigmoid	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
NSB-top	100	1	100	1	100	1
CT-top	100	0.91	100	0.91	100	0.91
IP	100	1	100	1	100	1
MP	88.89	0.92	88.89	0.92	88.89	0.92
Total	96.67		96.67		96.67	

Idaho – deep soil

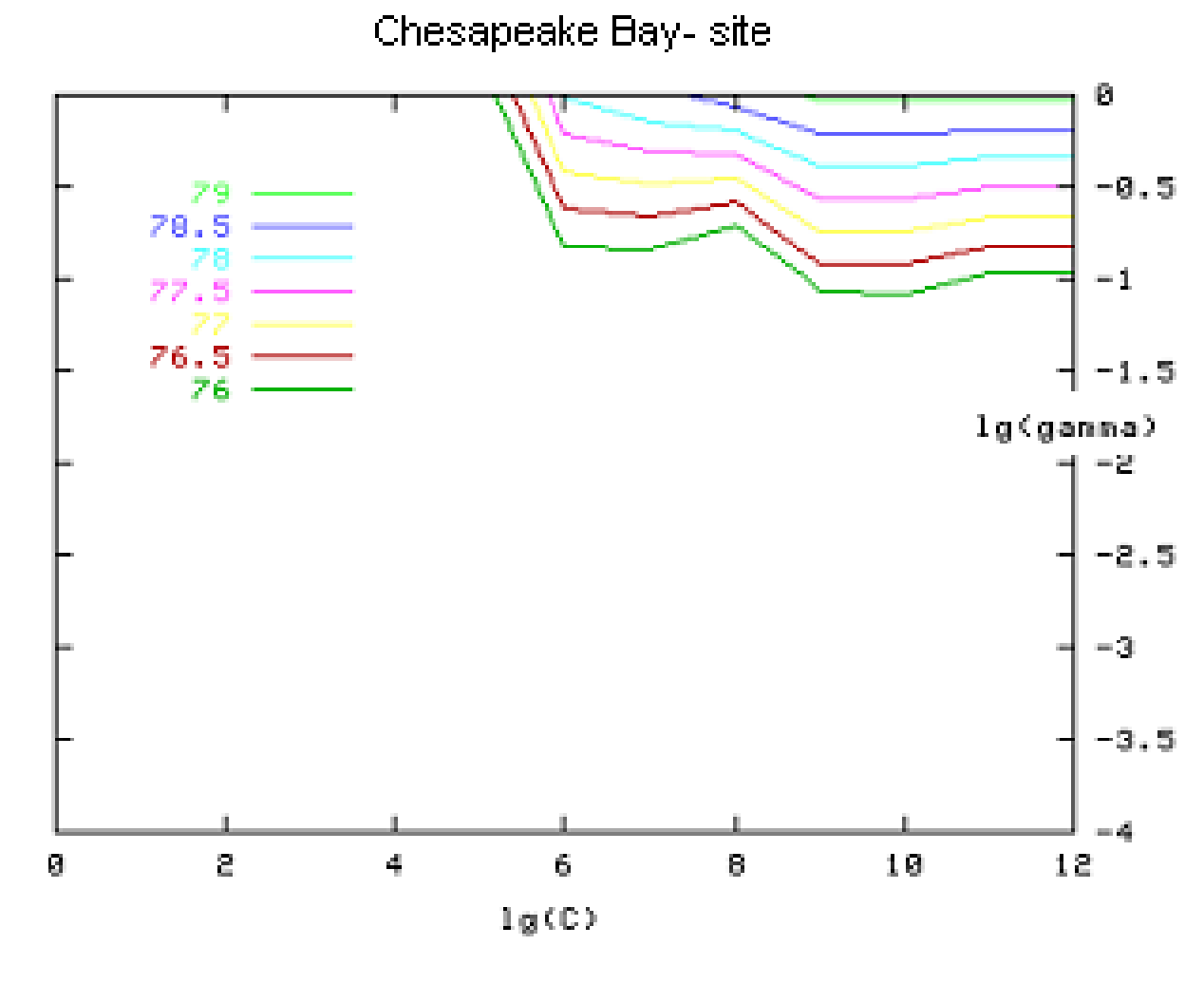
Location	Linear		RBF		Sigmoid	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
NSB-deep	100	0.85	100	0.85	100	0.85
CT-deep	100	0.78	100	0.78	100	0.78
IP	100	0.92	100	0.92	100	0.92
MP	88.89	0.85	88.89	0.85	88.89	0.85
Total	96.67		96.67		96.67	

Chesapeake Bay - site

Site	Number of samples	Accuracy (%)		
		Linear	RBF	Sigmoid
CP	23	0	0	0
CS	60	100	100	100
HD	52	0	0	0
HI	38	0	0	0
HW	42	0	0	0
OC	23	0	0	0
OM	9	0	0	0
RB	30	0	0	0
UP	5	0	0	0
Total	282	21.27	21.27	21.27

SVM

Model Selection



Chesapeake Bay - Site

Site	Number of samples	RBF*	
		Accuracy (%)	MCC
CP	23	78.26	0.80
CS	60	88.33	0.84
HD	52	90.38	0.88
HI	38	86.84	0.81
HW	42	80.95	0.83
OC	23	82.61	0.74
OM	9	55.56	0.48
RB	30	73.33	0.75
UP	5	60.00	0.67
Total	282	82.97	

Chesapeake Bay - time

Time	Number of samples	RBF*	
		Accuracy (%)	MCC
Sept 99	76	85.53	0.78
Dec 99	58	77.59	0.68
Mar 00	33	75.76	0.71
May 00	15	73.33	0.85
July 00	15	100.00	0.91
Nov 00	35	82.86	0.82
Feb 00	50	78.00	0.80
Total	282	81.20	

Conclusion

- Performed well using a simple kernel for Idaho data
- Model optimization necessary for unbalanced or large multiclassification
- Very effective for this ecogenomics problem
- This is the preliminary SVM study for ALH research

Future Work

- New kernel development
- Automation & GUI
- Interpretation
- Dealing with partial data

Discussions

- Cluster according to the soil management practices employed
- Devise broad classes of signatures of microbial communities
- Classify them according to the chosen classification scheme
- ALH profiles of a large collection of known samples of soils from different regions can be used to train a program

Acknowledgements

- Dr. Giri Narasimhan
- Dr. DeEtta Mills
- Dr. Chih-Jen Lin (LibSVM)