Performance Analysis of Alternative Combinations of Classification and Clustering Algorithms with Applications to Microbial Community Profiling

Rebecca Le

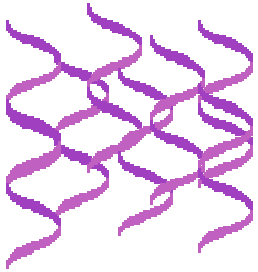
Daniel Jeske, Ph.D

Applied Statistics

University of California, Riverside

Background

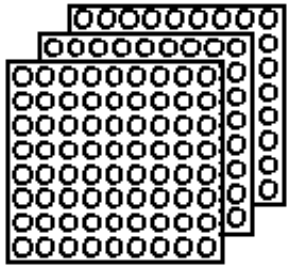
Long, unknown rRNA gene sequences

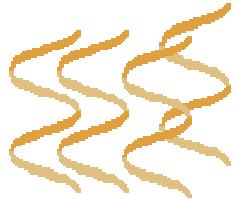


Extract DNA
PCR Amplify rRNA
Construct Clone Library

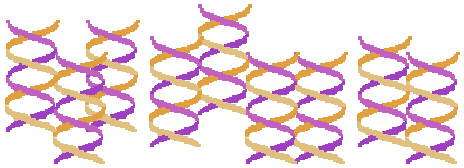
Samples (mouse guts, plant tissues, soil,…)

Array rRNA Genes





Hybridize Arrays with Probes



Signal Intensities

Classification

Fingerprints

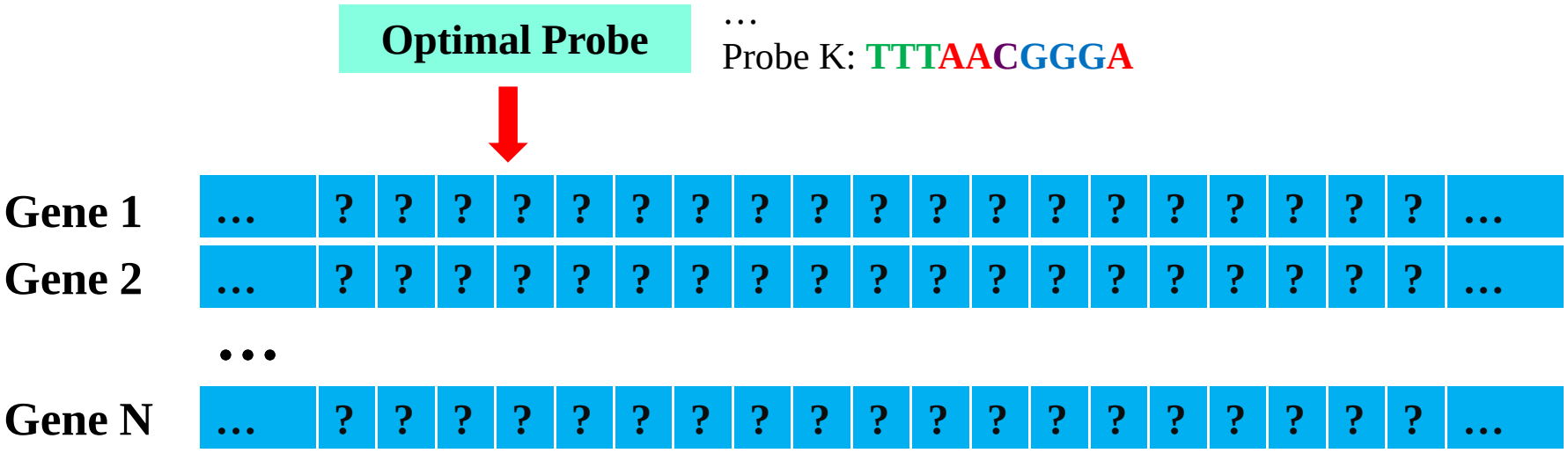
Clustering

Taxonomic Tree

Identify rRNA Genes

Data Collection

Probe1: AAAGGTTTTT
Probe 2: CTTCCCCAAT
Probe 3: GGGGAATTCC
…
Probe K: TTTAACGGGA



Output Intensity Measurements

Gene 1	0.23	2.99	1.02	2.88	…	1.11
Gene 2	1.22	2.44	0.99	0.21	…	1.33
…	…	…	…	…	…	…
Gene N	3.05	0.01	3.44	2.56	…	0.78

K times

Neutral Zone Classifiers (NZC) 2-NZC & 2-BC

Costs of Misclassification Errors

True Class Label	Predicted Class Label		
	0	1	N
0	0	C_1	C_2
1	C_1	0	C_2

- 2-NZC can be expressed as:

$$\hat{C}(y;L)=\begin{cases}0 & if \quad p_0(y)>1/2+L/2\\ 1 & if \quad p_0(y)<1/2-L/2\\ N & otherwise\end{cases}$$

where $p_i(y)=P(C=i|Y=y)=\pi_i f_i(y)/\sum_i \pi_i f_i(y)$ $\quad i=0,1$ and $f_i(y)=f_{Y|C=i}(y)$

Find the limit L:

- When $\rho=C_1/C_2$ is known, find the limit L by minimizing:

$$E[Error\ \ Cost\ ]\propto f(L)=\pi_0[\rho P(\hat{C}=1|C=0)+P(\hat{C}=N|C=0)]$$
$$+\pi_1[\rho P(\hat{C}=0|C=1)+P(\hat{C}=N|C=1)]$$

- When ρ is not known, see the reference for details.

- 2-BC: the mechanism is same as 2-NZC, but posterior probabilities are putting in the fingerprints.

Jeske, D.R., Liu, Z., Bent, E., and Borneman, J. (2007). Classification rules that include neutral zones and their application to microbial community profiling. Communication in Statistics – Theory and Methods 36(10): 1965-1980.

Aggregating Neutral Zone Classifier (ANZC)
Bayes Classifier (BC) and Aggregating Bayes Classifier (ABC)

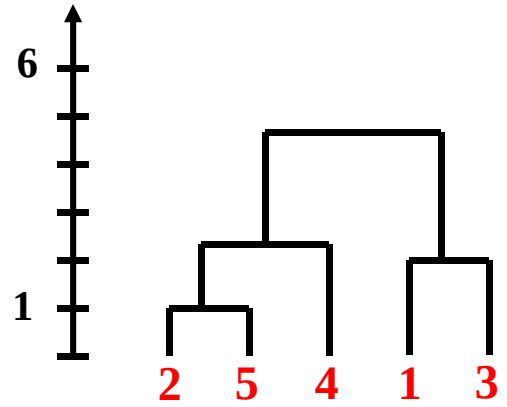
- "Bagging" (**B**ootstrap **agg**regat**ing**) algorithm:
  - Goal: improve the accuracy and stability of the predictions.
  - Consider an original training dataset $D = \{(C_i, y_i), i=1,\ldots,N\}$
  - Form $\{\hat{C}(y,D^*)\}$ where $\{D^*\}$ are bootstrap samples from D.
  - Aggregating classifier:
    $\hat{C}_A(y)$ = {majority voting if *C's* are class labels}

- ANZC: $\hat{C}_{ANZC}(y)$ = {majority voting if *C's* are class labels}
- ABC: $\hat{C}_{ABC}(y)$ = {averaging y_i across $\hat{C}_i(y_i,D^*)$}

Unweighted Pair Group Method Average (UPGMA)

Algorithm (for continuous measurements):

1. Compute an NxN symmetric distance matrix of N genes.
2. Search the distance matrix for the nearest (most similar) pair of clusters.
3. Merge these two clusters and update the distance matrix.
4. Repeat (2) and (3) a total of (N-1) times.



Sneath, P.H.A & Sokal, R.R. (1973). Numerical taxonomy - the principles and practice of numerical classification. Freeman, San Francisco.

Greedy Clique Partition (GCP)

- Use graph theory to find clusters; treat N as a missing value.
- Terminology: Consider a set of 0/1/N fingerprints $F=\{f_1, f_2, \ldots, f_N\}$.
  - f_i and f_j are resolved: $f_i = (0,1,1,0,1)$ and $f_j = (0,1,1,0,1)$
  - f_i and f_j are compatible: $f_i = (1,1,N,0,1,1)$ and $f_j = (1,1,0,0,1,1)$

- Vertices: fingerprints.
- Edges: representatives of compatible between fingerprints.
- Clique: portion of a graph where every two vertices are connected.
- Maximum clique: Clique contains a largest numbers of vertices.
- Unique maximum clique: has all compatible neighbors.

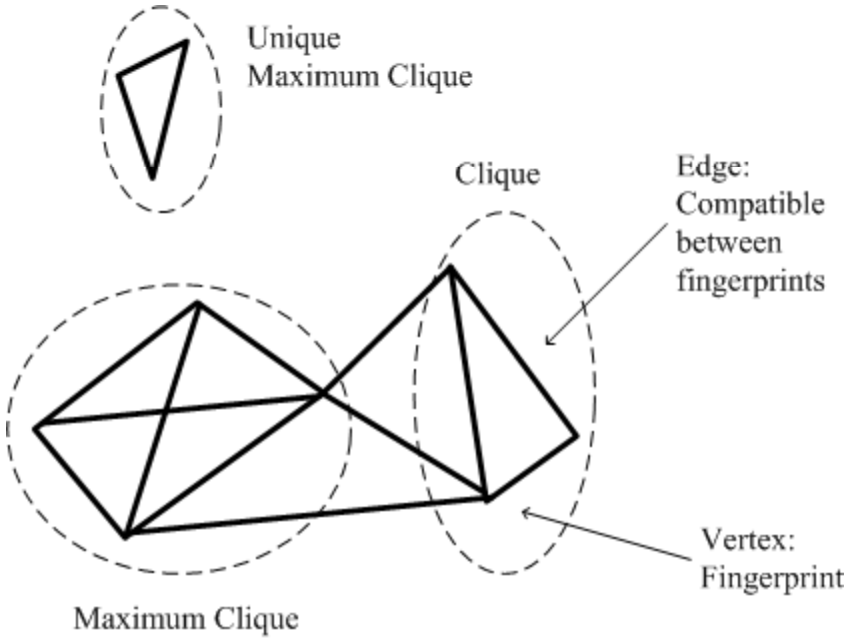


Figure 1. Illustration of GCP Algorithm Terminology

Figueroa, A., Borneman, J., and Jiang, T. (2004). Clustering binary fingerprint vectors with missing values for DNA array data analysis. Journal of Computational Biology 11(5): 887-901.

GCP Algorithm

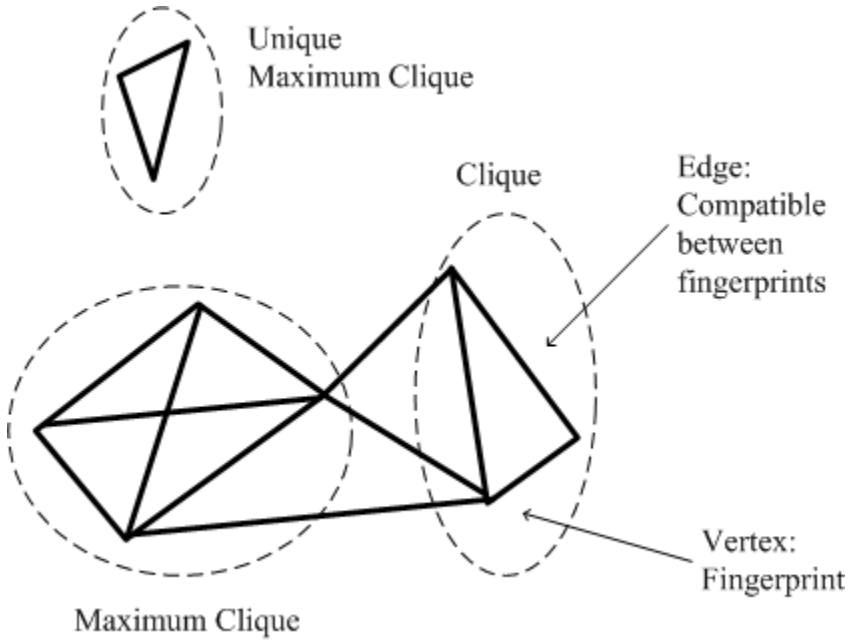
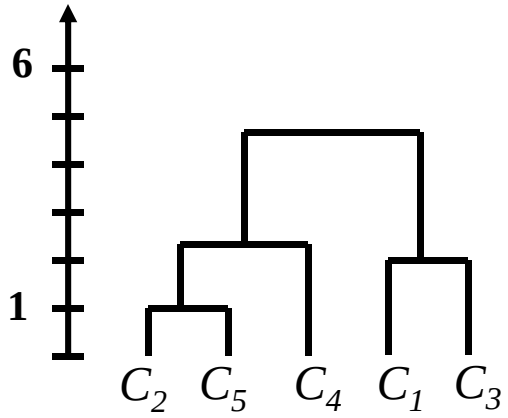


Figure 1. Illustration of GCP Algorithm Terminology

1. Search and remove a unique maximal clique C_u from the graph; add C_u to C.
2. Repeat step 1 until no more unique maximal clique left.
3. Search and remove a maximum clique C_m from the graph; add C_m to C.
4. Repeat (1) to (3) until all fingerprints are added to C.

- **Outcome 1**: A set of clusters $C=\{C_1, C_2, \ldots, C_m\}$ where C_i is a set of mutually compatible fingerprints.
- **Outcome 2**: Create representative fingerprints for each cluster C_i and build a cluster dendrogram.



A cluster dendrogram

Proposed Joint Classification and Clustering Methods

- Method 1: 2-BC+UPGMA
- Method 2: 2-ABC+UGPMA
- Method 3: 2-NZC+GCP
- Method 4: 2-ANZC+GCP

Recall: 2-BC: Two-Classes Model Bayes Classifier

2-ABC: 2-BC employing “bagging”

2-NZC: Two-Classes Model Neutral Zone Classifier

2-ANZC: 2-NZC employing “bagging”

GCP: Greedy Clique Partition

UPGMA: Unweighted Pair Group Method Average

Between-Dendrogram Similarity Metrics

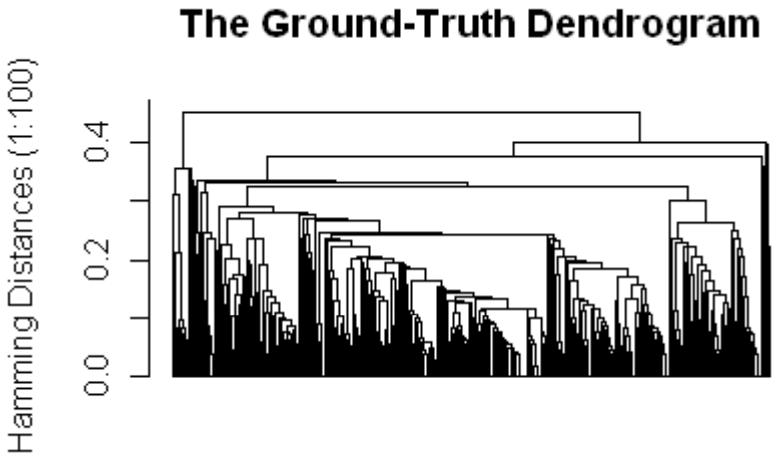
- Rand Index: $RI = (a+d) / (a+b+c+d)$
- Jaccard Index : $JI = a / (a+b+c)$
- Fowlkes & Mallows Index : $FM = a / sqrt((a+b)(a+c))$
- Adjusted Rand Index: $ARI = [RI – E(RI)] / [max(RI) – RI]$

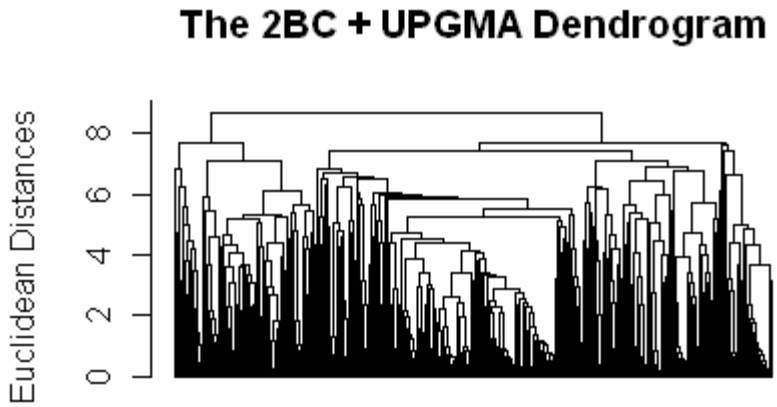
The most sensitive metric for detecting similarity between two dendrograms.

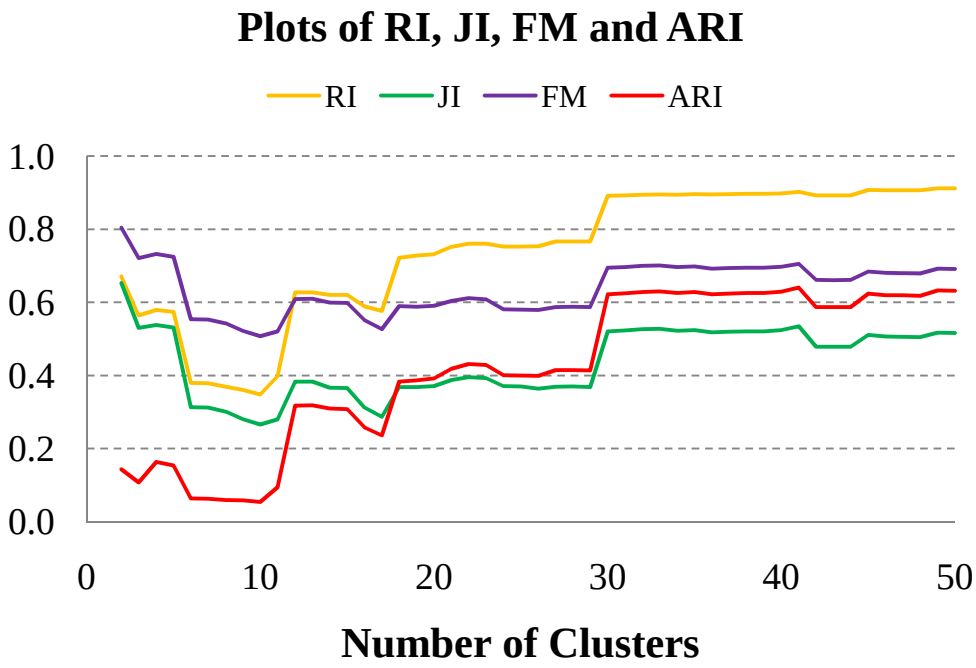
- If ARI > 0.90: Excellent consistency.
- If ARI > 0.80: Good consistency.
- If ARI > 0.65: Moderate consistency.
- If ARI < 0.65: Poor consistency.

1. Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66: 846-850.
2. Everitt, E. (1993). Cluster analysis. Edward Arnold, London, 3rd edition (presented 1980).
3. Fowlkes, E.B., & Mallows, C.L. (1983). A method for comparing two hierarchical clustering. Journal of the American Statistical Association 78: 553-569.
4. Hubert, L. & Arabie, P. (1985). Comparing partitions. Journal of Classification 2(1): 193-218.
5. Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand Index. Psychological Methods 9(3): 386-396.

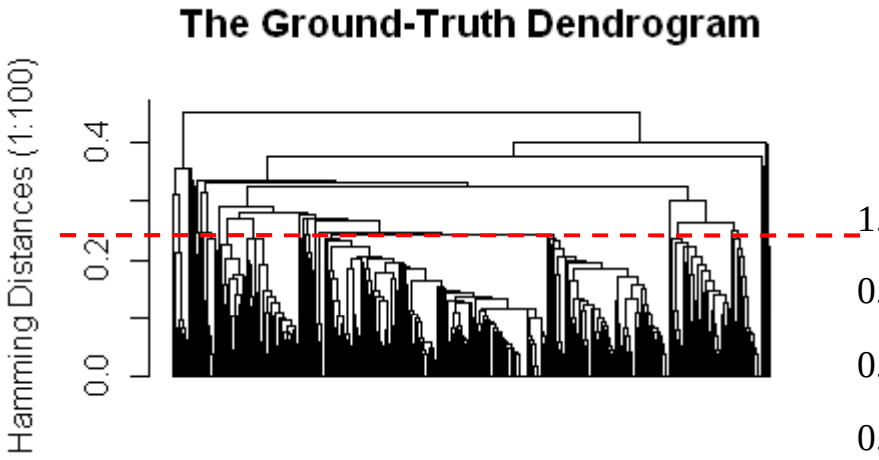
Result: 2BC+UPGMA and the Ground-Truth

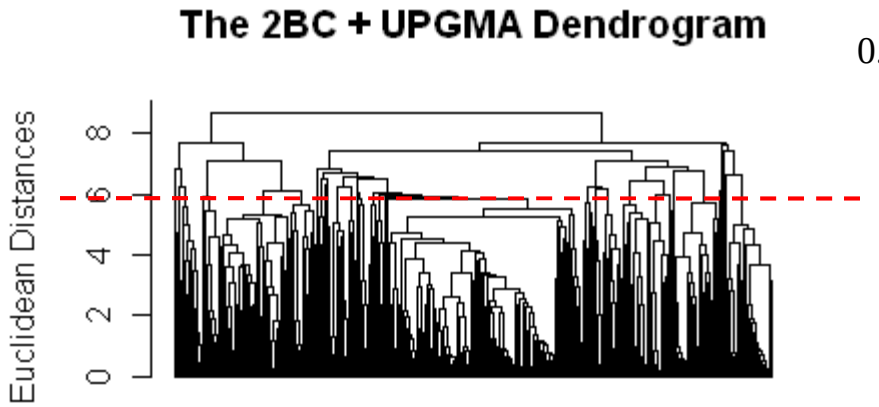


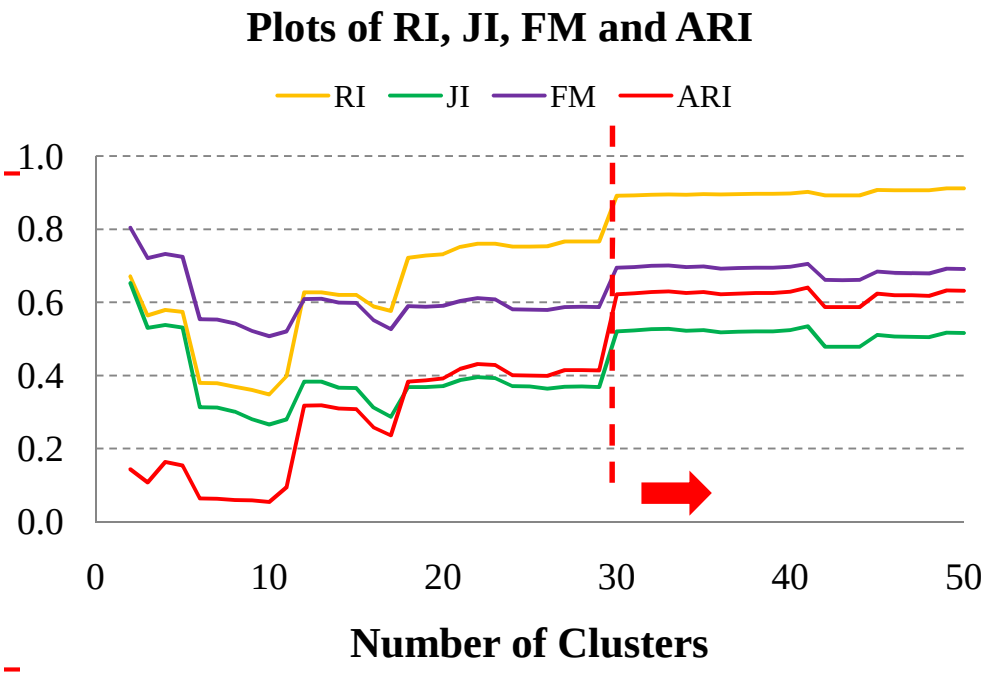




Result: 2BC+UPGMA and the Ground-Truth





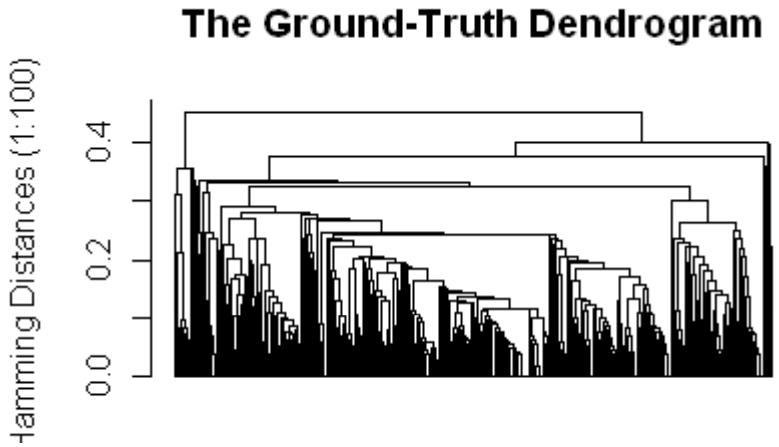


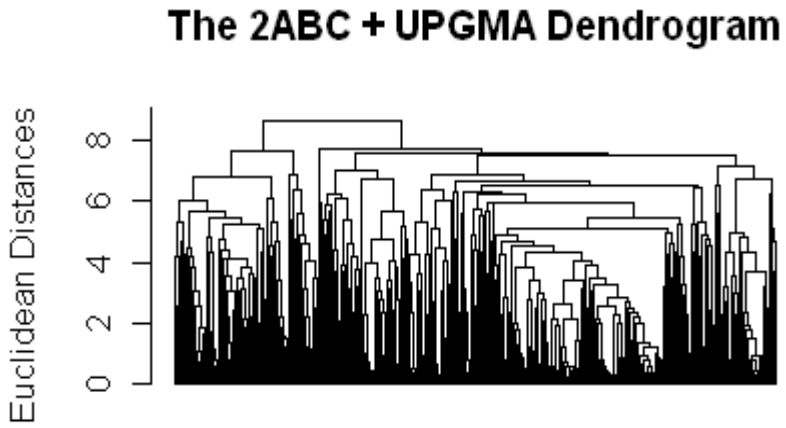
When number of clusters $\geq$ 30:

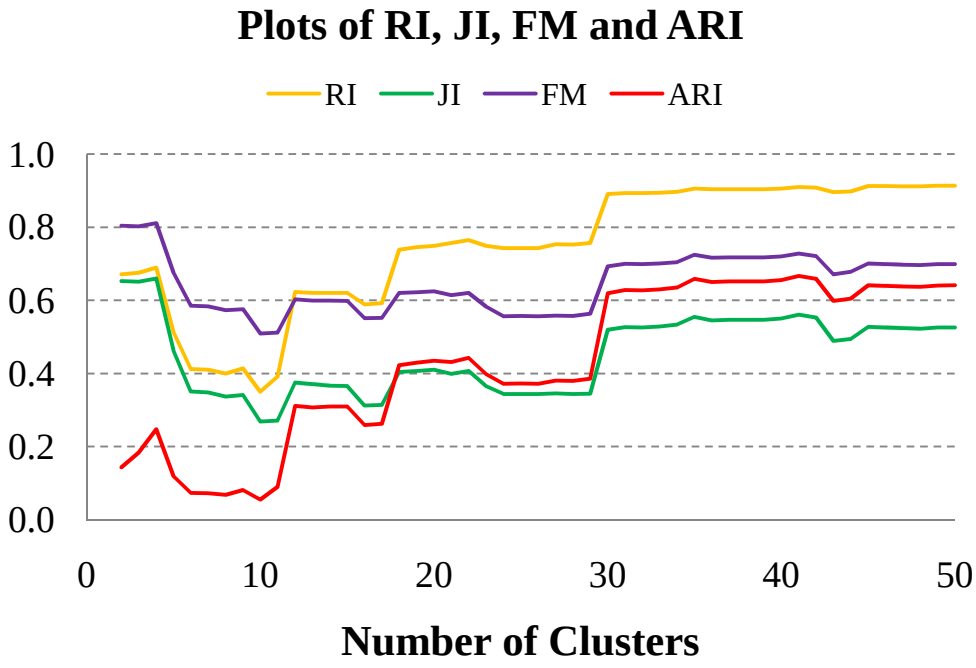
ARI $\geq$ 0.60 ($\approx$0.65)

→2BC+UPGMA shows evidences of a moderate agreement with the ground-truth dendrogram.

Result: 2ABC+UPGMA and the Ground-Truth

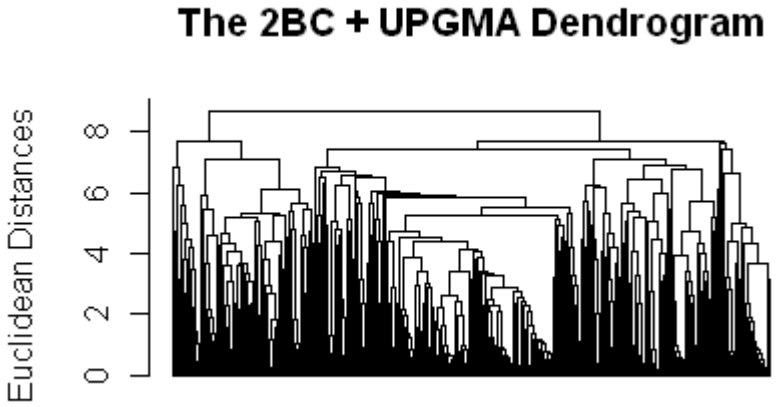


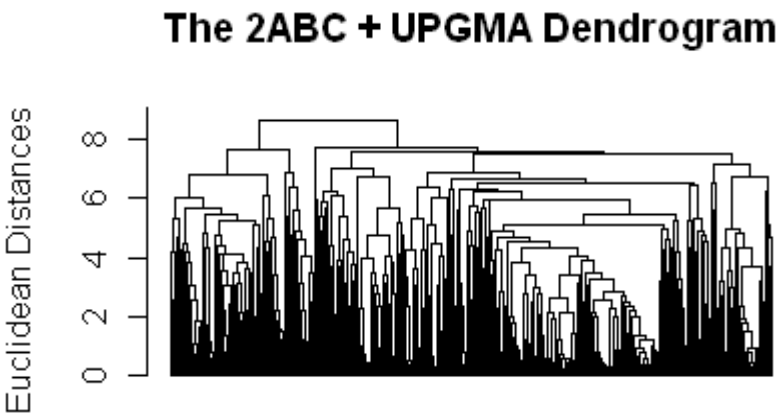


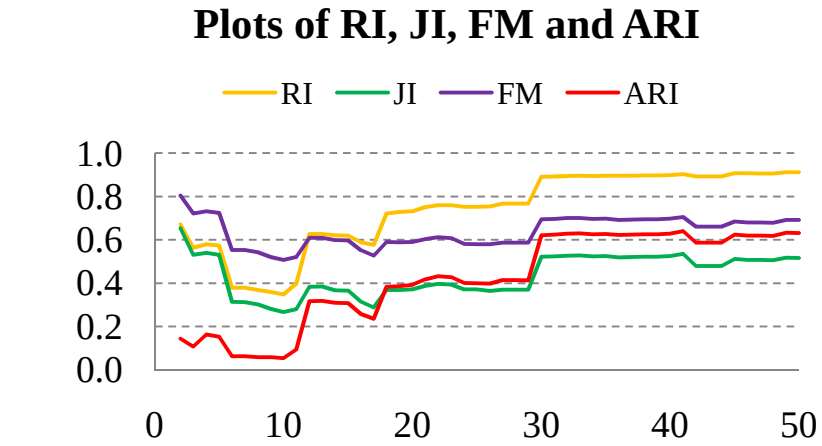


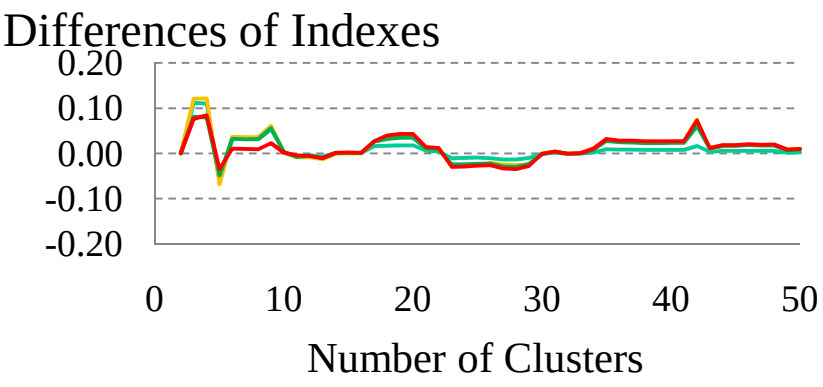
The best result occurs when employing 40 bootstrap samples.

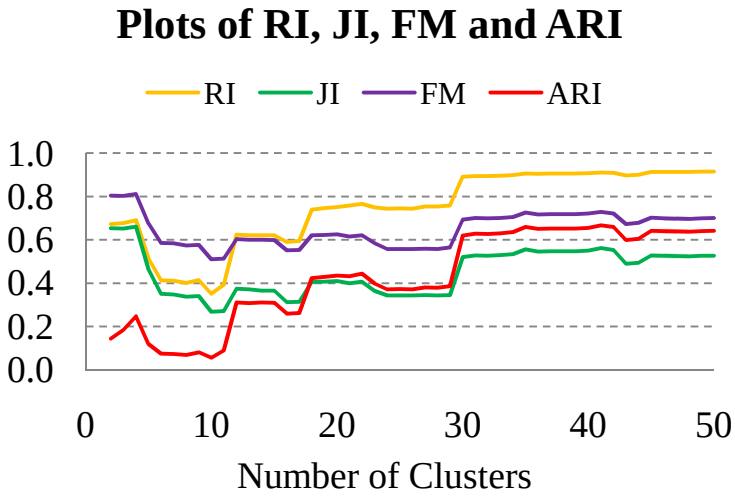
Is 2ABC+UPGMA better than 2BC+UPGMA?





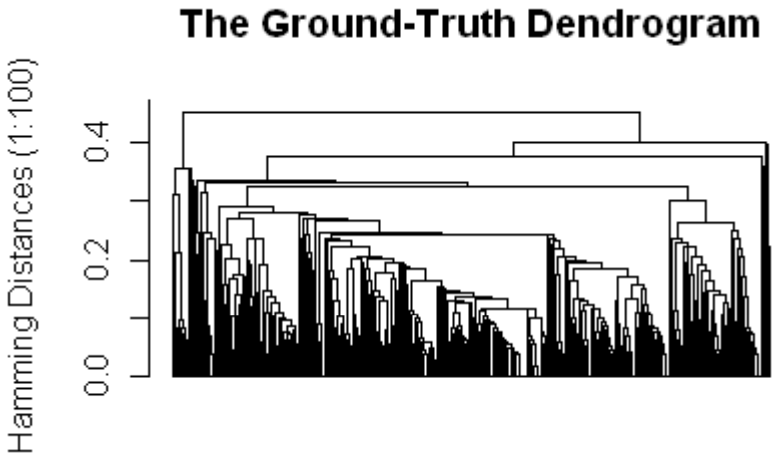


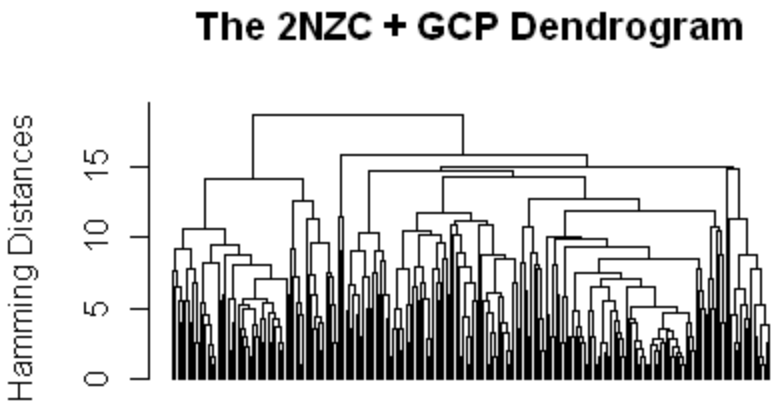


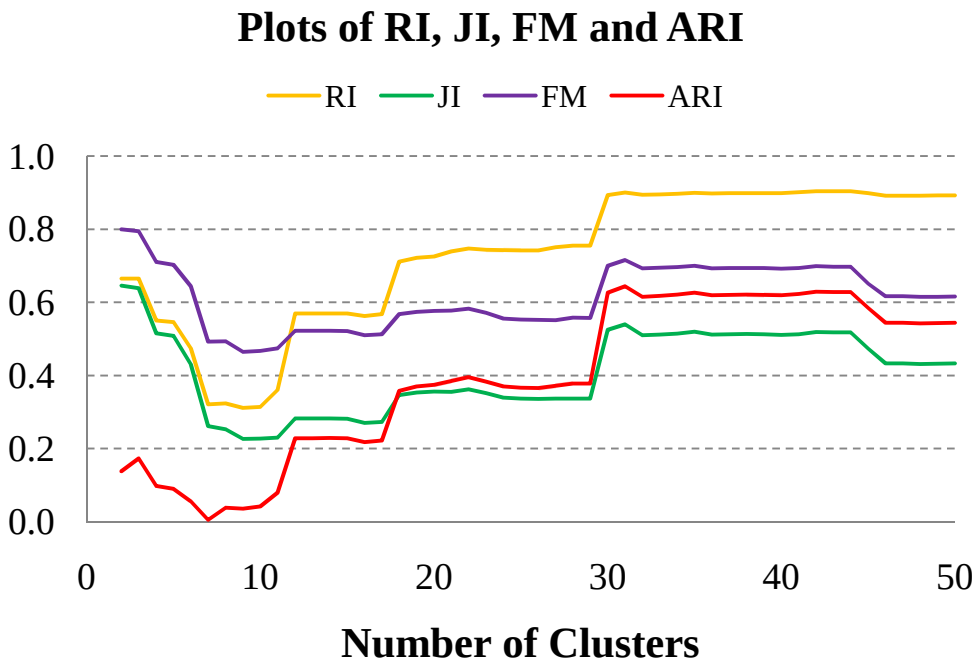


There are no significant and consistent improvements of 2ABC+UPGMA over 2BC+UPGMA.

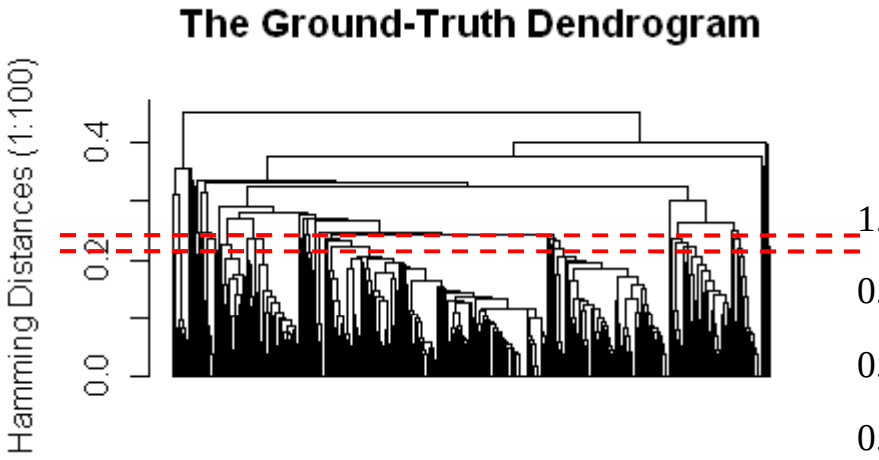
Result: 2NZC+GCP and the Ground-Truth

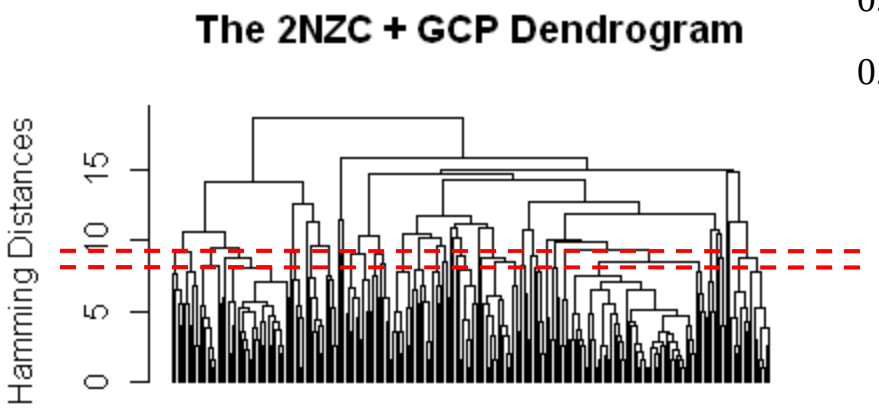


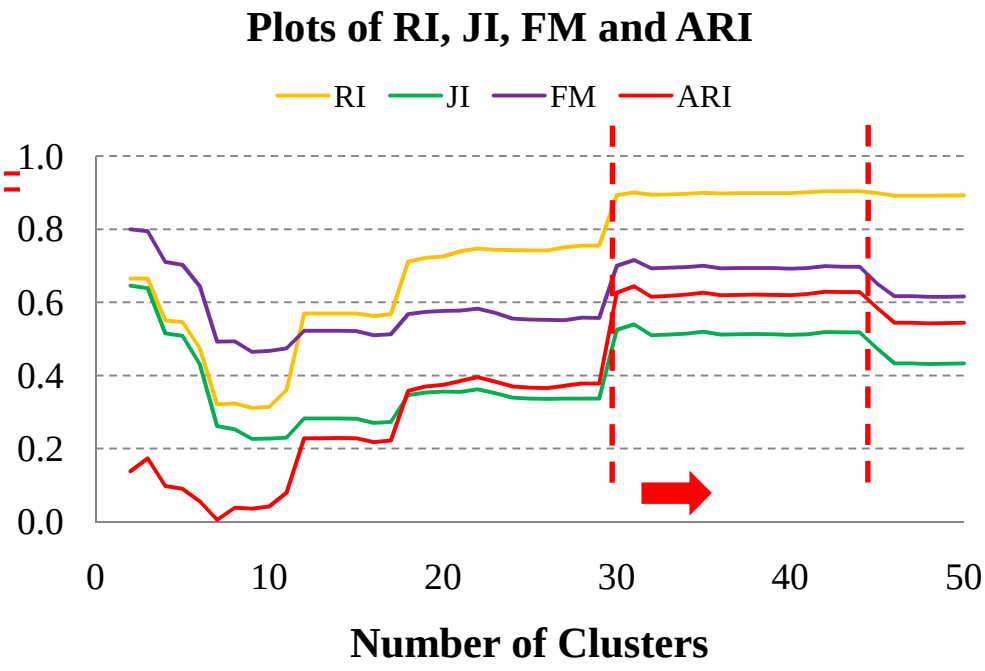




Result: 2NZC+GCP and the Ground-Truth





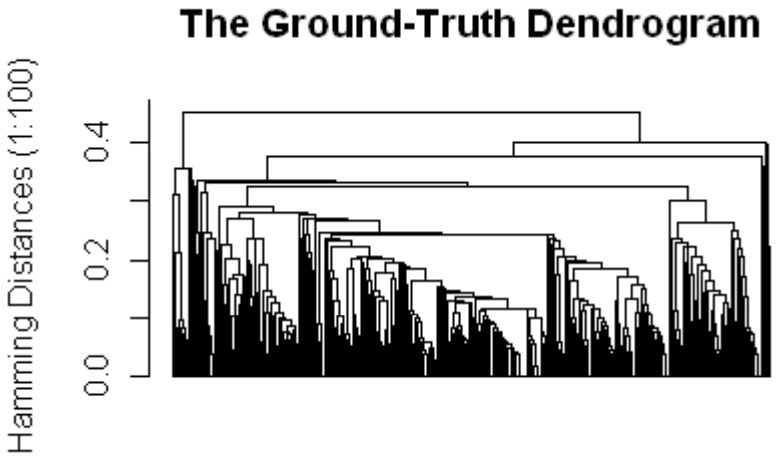


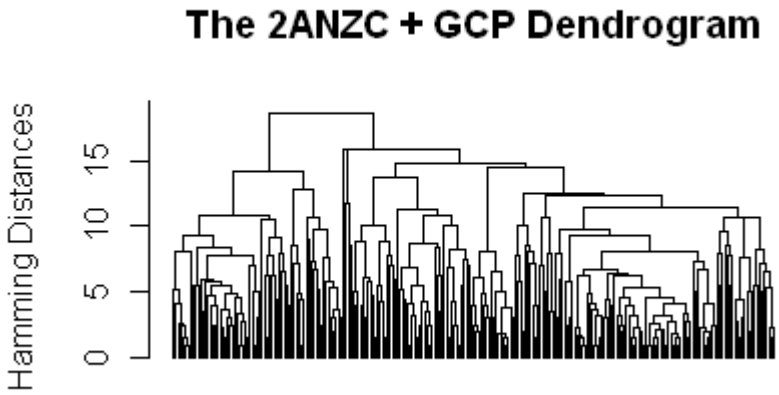
When number of clusters from 30 to 45:

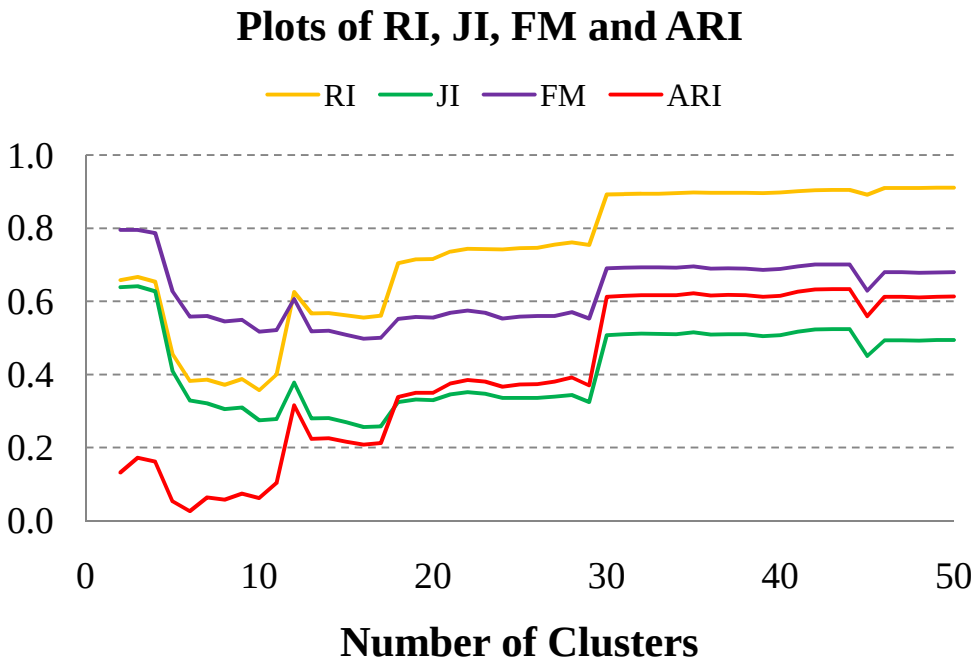
ARI ≥ 0.60 (≈0.65)

→2NZC+GCP shows evidences of a moderate agreement with the ground-truth dendrogram.

Result: 2ANZC+GCP and the Ground-Truth

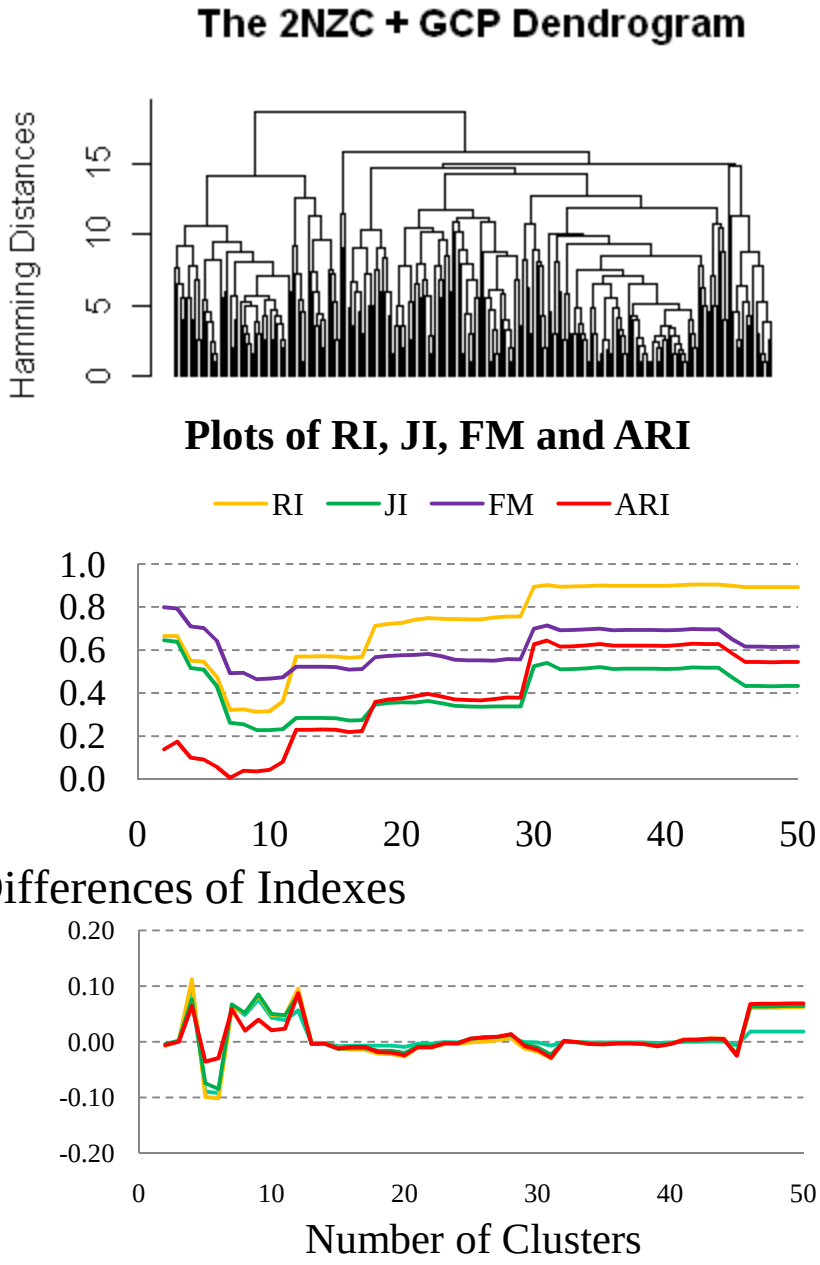


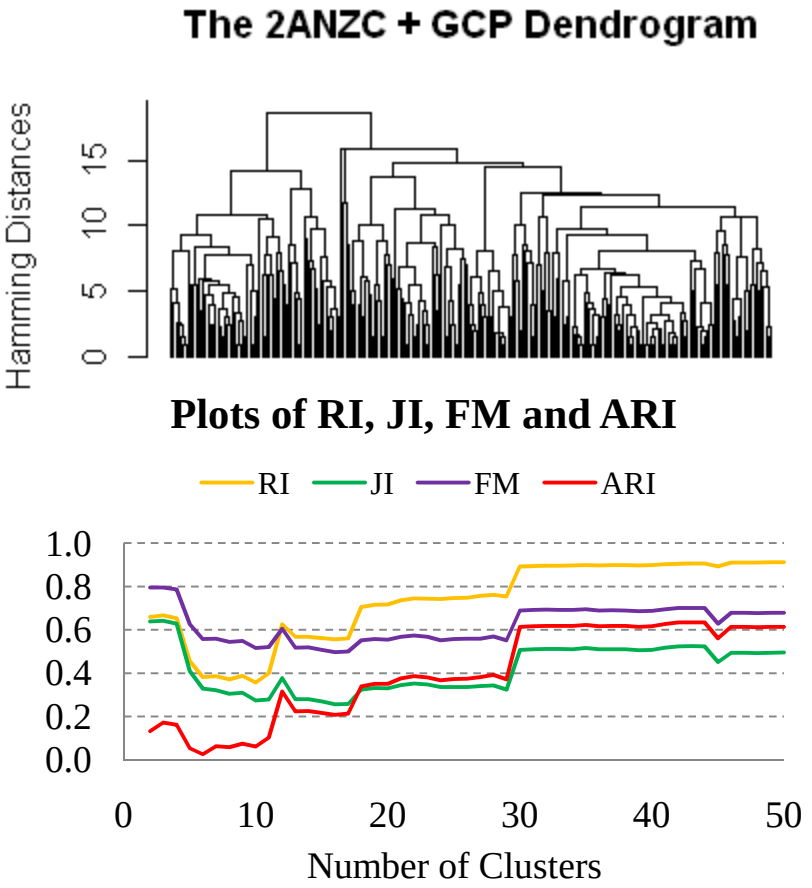




The best result occurs when employing 25 bootstrap samples.

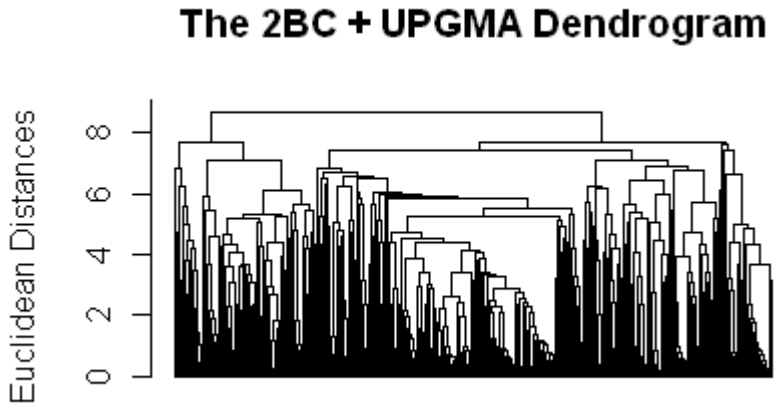
Is 2ANZC+GCP better than 2NZC+GCP?

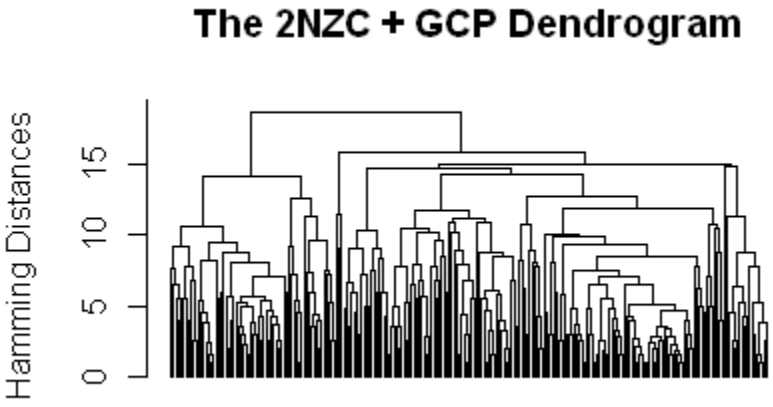




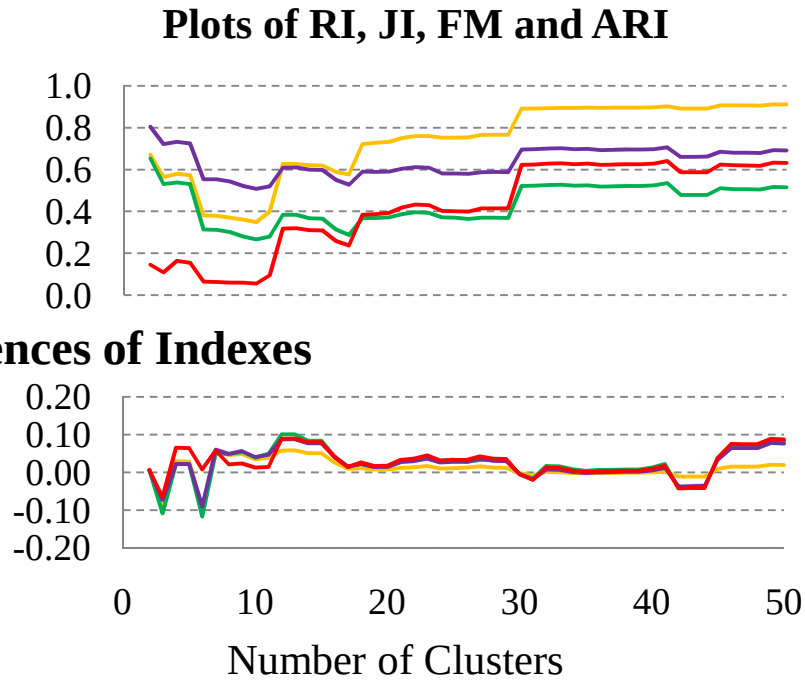
There are no significant and consistent improvements of 2ANZC+GCP over 2ANZC+GCP.

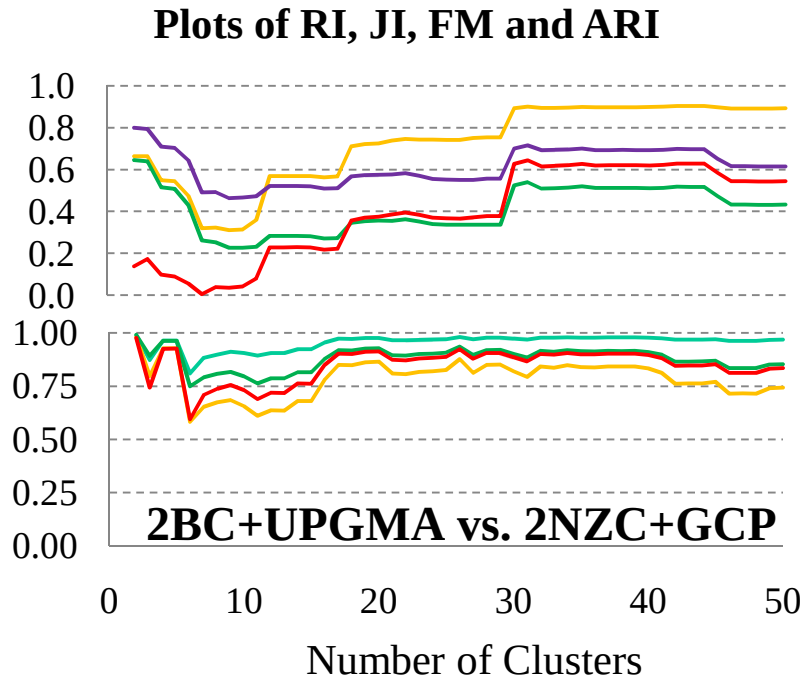
2BC+UPGMA vs. 2NZC+GCP





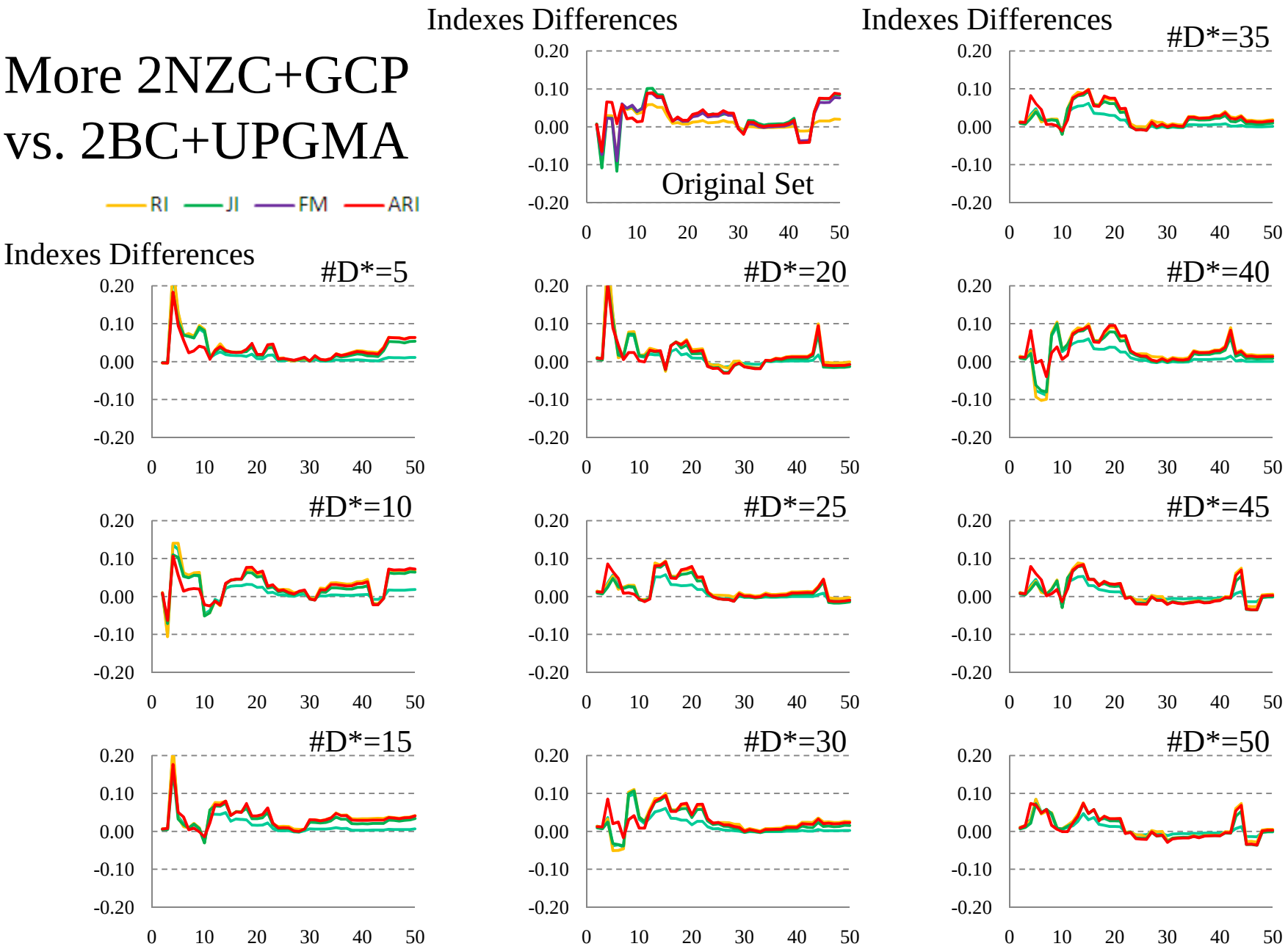
Compare with the ground-truth.





2BC+UPGMA and 2NZC+GCP are almost identical.

More 2NZC+GCP vs. 2BC+UPGMA



ARI of 2BC+UPGMA are consistently higher than 2NZC+GCP across all comparisons with the ground-truth dendrogram → higher agreements with the ground-truth results.

Conclusions

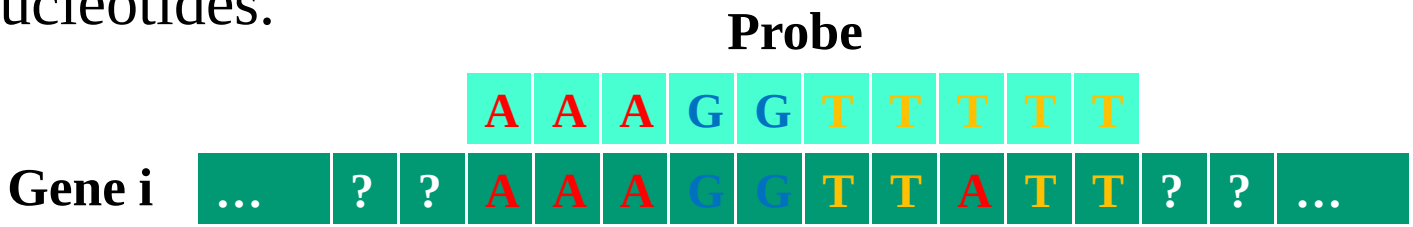
2BC+UPGMA and 2NZC+GCP methods show a promising agreement with the ground-truth dendrogram.

2BC+UPGMA and 2NZC+GCP methods should be used to generate clustering dendrograms instead of 2ABC+UPGMA and 2ANZC+GCP in term of a time-effectiveness.

2BC+UPGMA method has more advantages than 2NZC+GCP in terms of simple algorithm and a little more identical to the ground-truth dendrogram.

Future Work

- Consider partial binding allowing one unmatched pair of nucleotides.



- Consider utilizing association between intensity measurements across probes.

Thank you.