A comparative study of unsupervised machine learning methods for mircoRNA sequence-based clustering

Samuel Acosta-Melgarejo
Supervisor: Sam Griffiths-Jones

Faculty of Biology, Medicine and Health, University of Manchester

BIOL61230/Research Project 1

Background

MicroRNAs

Background

MicroRNAs

- Small non-coding RNA molecules of about 22 nucleotides, found in animals, plants and some viruses.

Background

MicroRNAs

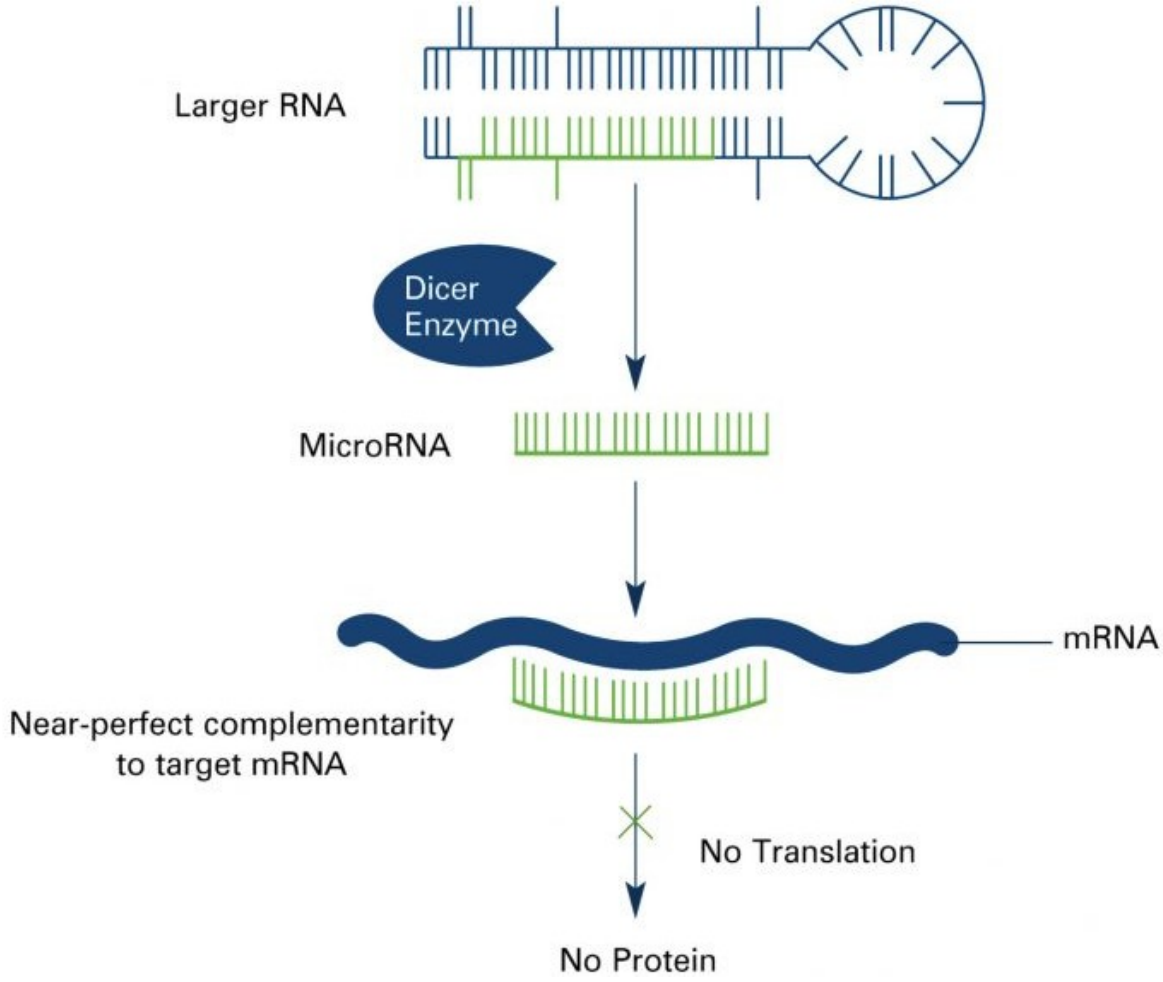
- Small non-coding RNA molecules of about 22 nucleotides, found in animals, plants and some viruses.
- Important post-transcriptional functions in gene expression, (developmental timing, cell death, cell proliferation...).

Background

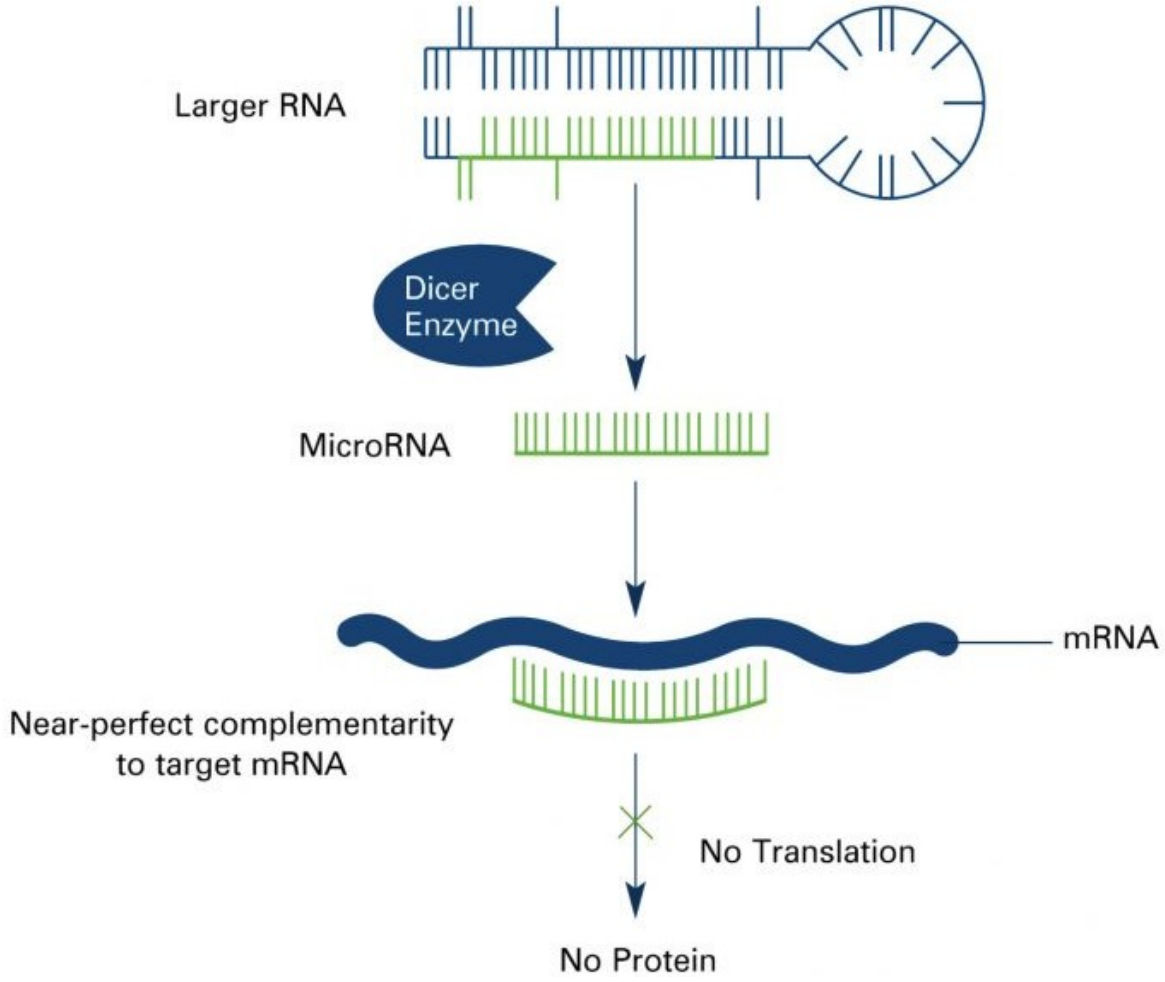
MicroRNAs

- Small non-coding RNA molecules of about 22 nucleotides, found in animals, plants and some viruses.
- Important post-transcriptional functions in gene expression, (developmental timing, cell death, cell proliferation...).
- Research and annotation centralised in the miRBase database.

Background

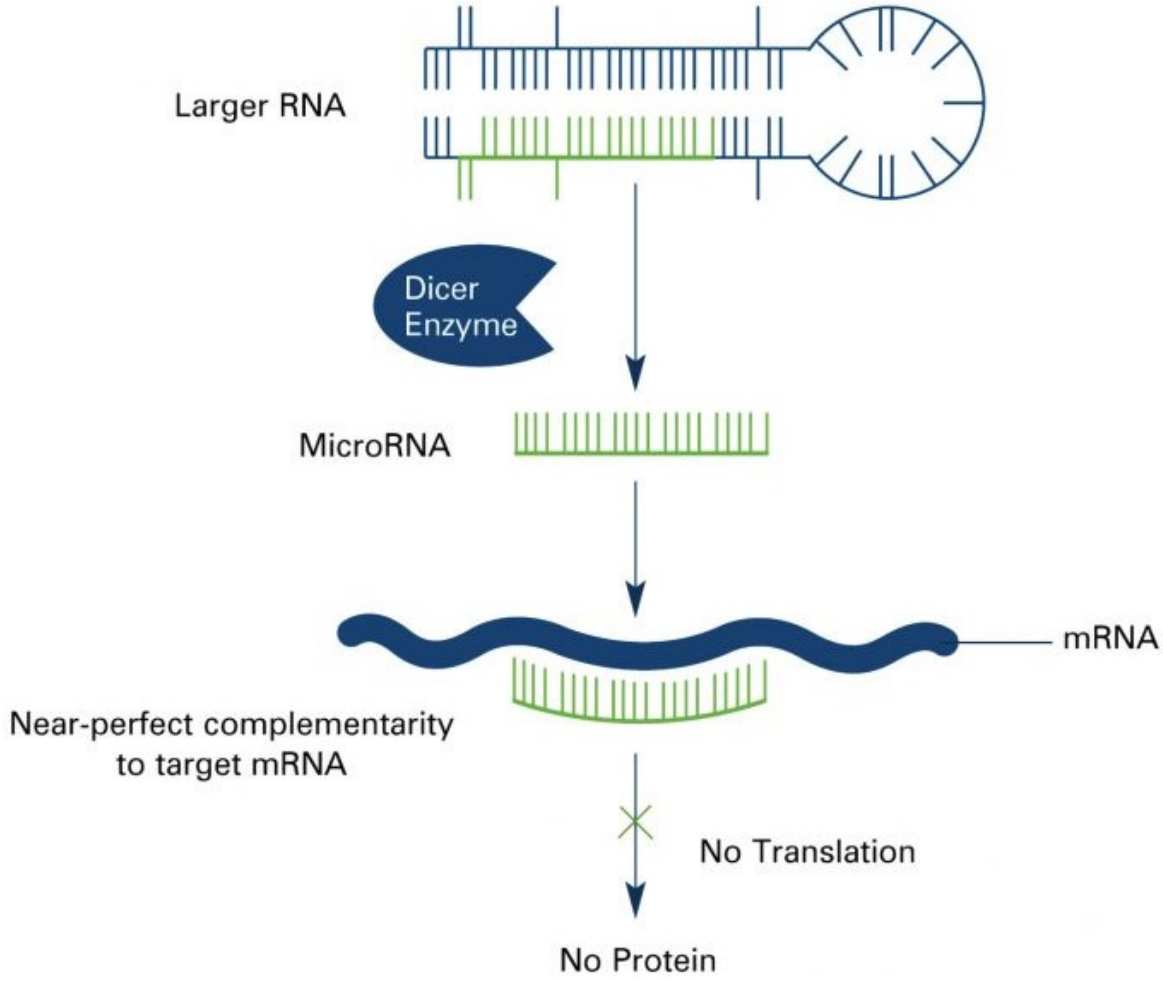

Larger RNA
Dicer
Enzyme
MicroRNA
mRNA
Near-perfect complementarity
to target mRNA
No Translation
No Protein

Background



- Excised by enzymes from longer precursors.

Background



- Excised by enzymes from longer precursors.
- Precursors folded into hairpin-like secondary structures.

Background

MiRNA families

- Groups of pre-miRNAs based on different criteria (similar ancestry, secondary structure conservation, seed-target relations...)
- An important aspect used is the analysis of sequence similarity.

Background

MiRNA families

- Groups of pre-miRNAs based on different criteria (similar ancestry, secondary structure conservation, seed-target relations...)
- An important aspect used is the analysis of sequence similarity.

Why are they important?

- Valuable information about biological functions.
- Amount of information about different miRNAs is variable; families useful for hypothesising characteristics of less known miRNAs.
- New miRNAs discovered at a fast rate; even low confidence families still useful for researchers (available way before biological validation).

Introduction

Project aims

- Create a computational tool that allows to automatically detect miRNA families from miRBase database (manual process).
- Establish quality and significance of detected families.
- Compare different computational approaches and determine the most suitable.

Introduction

Project aims

- Create a computational tool that allows to automatically detect miRNA families from miRBase database (manual process).
- Establish quality and significance of detected families.
- Compare different computational approaches and determine the most suitable.

Requirements

- It was desirable that the tool could predict without human intervention $\rightarrow$ unsupervised machine learning
- Complete miRBase dataset, so results could be objectively compared to real families in miRBase $\rightarrow$ algorithms able to cluster large datasets

Methods: Algorithms

Unsupervised machine learning: Identifies groups of elements based on a similarity measure (obtained by embedded vectors/all-to-all BLAST).

1. Centroid-based clustering (*k-means++*, ClustalΩ impl.)

Methods: Algorithms

Unsupervised machine learning: Identifies groups of elements based on a similarity measure (obtained by embedded vectors/all-to-all BLAST).

1. Centroid-based clustering (*k-means++*, ClustalΩ impl.)
2. Stochastic graph-based clustering (MCL, ref. impl.)

Methods: Algorithms

Unsupervised machine learning: Identifies groups of elements based on a similarity measure (obtained by embedded vectors/all-to-all BLAST).

1. Centroid-based clustering (*k-means++*, ClustalΩ impl.)
2. Stochastic graph-based clustering (MCL, ref. impl.)
3. Density-based clustering (DBSCAN, scikit-learn impl.)

Methods: Statistical validation

1. Fowlkes-Mallows score

External similarity measure, expressed as the geometric mean of the pairwise precision and recall.

Definition

$$\mathrm{FMS} = \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}} \tag{1}$$

- TP = true positives
- FP = false posititves
- FN = false negatives

Methods: Statistical validation

2. Adjusted Rand index

Similar external similarity measure function, a version of the Rand index corrected for chance.

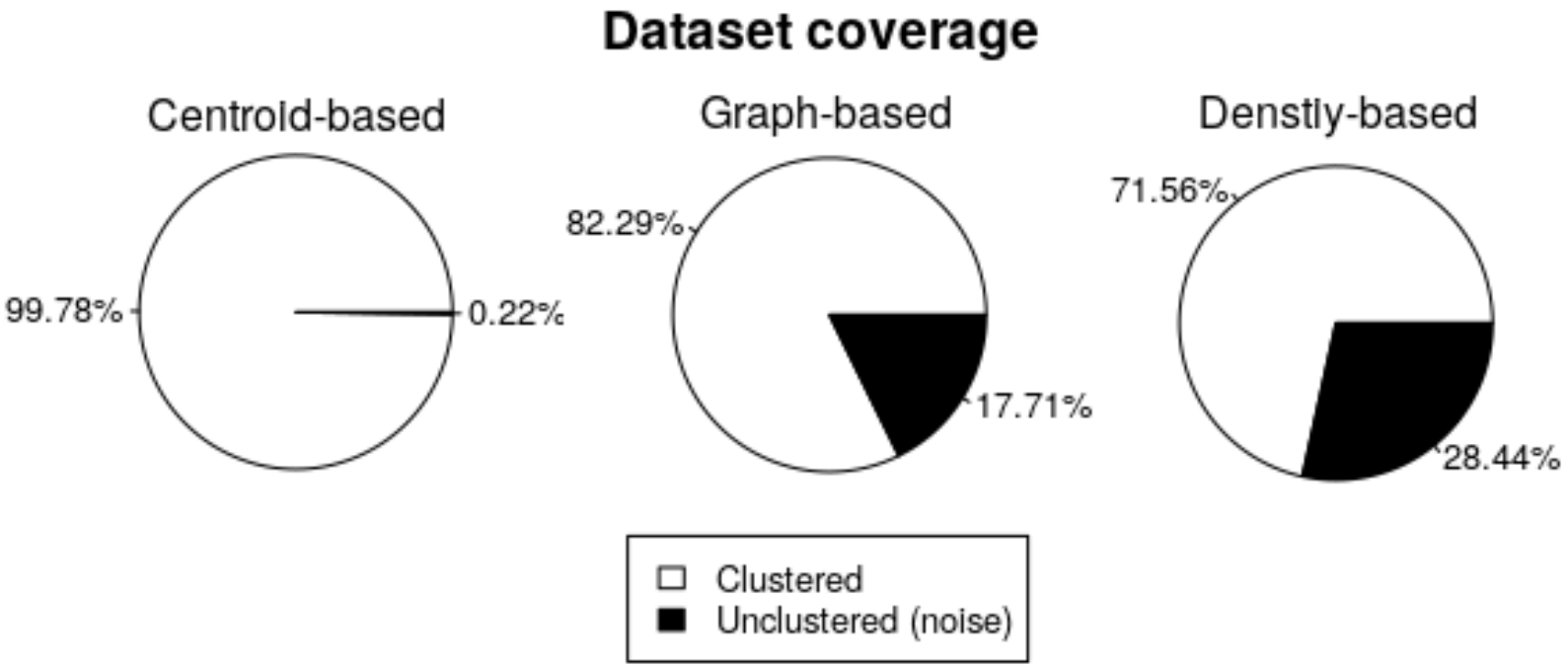
Definition

$$\mathrm{RI} = \frac{a+b}{C_2^{n_{samples}}} \quad ; \mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]} \tag{2}$$

- a = true positives
- b = true negatives
- $C_2^{n_{samples}}$ = total number of possible pairs
- $E[\mathrm{RI}]$ = expected index of random labellings

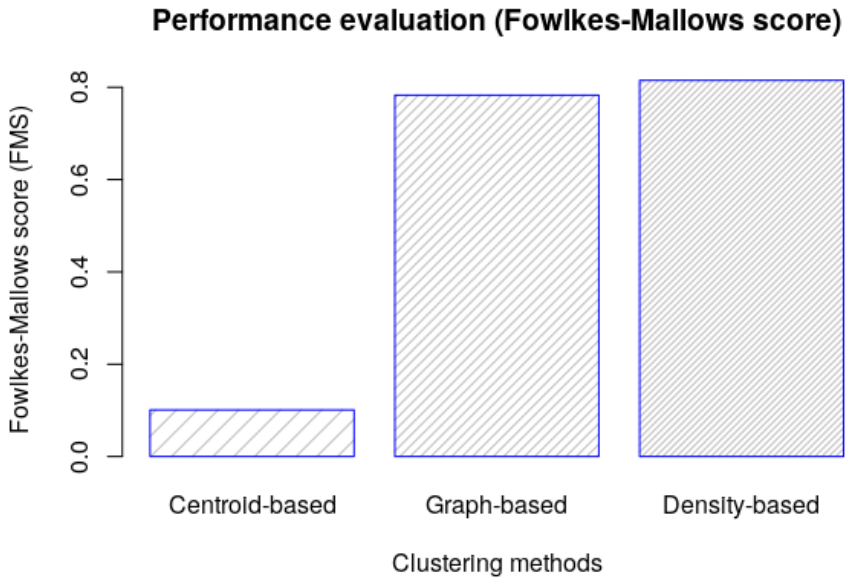
Results: Data coverage

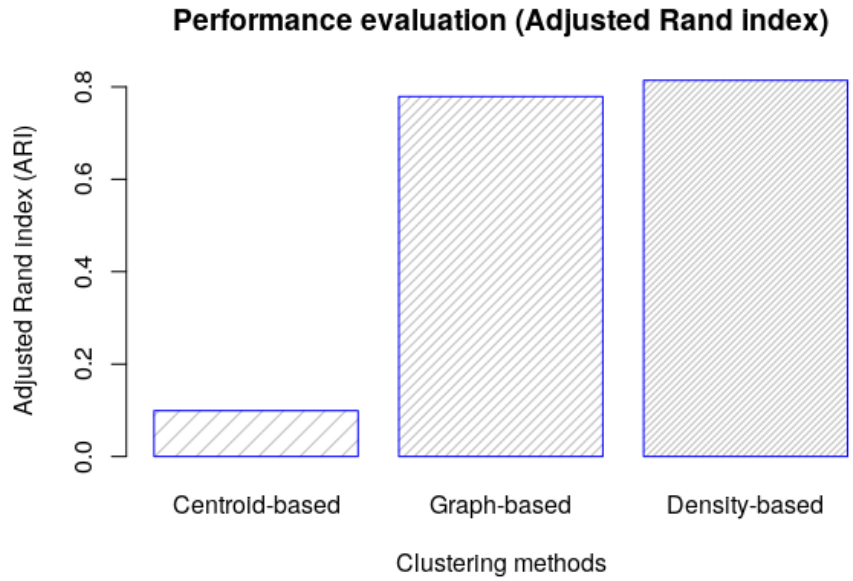
- Ref: 28645 miRNAs in 1983 families in miRBase (69.49% coverage).
- About 10% difference among graph and density-based algorithms.
- Centroid-based algorithm very different, clusters almost all the sequences.



Results: Quality assessment

- Similar patterns observed in FMS and ARI, the former tends to evaluate more positively.
- The centroid-based algorithm had a much lower score in both.
- The density-based algorithm had the best score in both measures, followed closely by the graph-based algorithm.





Results: Clustering overview

- Low performance of centroid-based algorithm explained by noise sensitivity and hard-wired cluster size threshold in implementation.
- Good performance of graph and density-based algorithms can be explained by their sparse-graph-oriented design.
- Stochastic graph-based algorithm has more coverage and clusters, implementation has better biological data support.
- ε value parameter in density-based algorithm increases noise detection and improves results.

Conclusions

- Unsupervised machine learning methods proven very useful for miRNA family detection in large datasets.
- Stochastic graph-based clustering, and specially, density-based clustering are the most suitable.
- The developed tool, miRNACluster, can be effectively used as an automated aid for miRNA family detection or to complement, adjust or improve the miRBase original family predictions.
- The tool is publicly available under an open source license at GitHub (www.github.com/samuacosta/miRNACluster)

Thank you.

samuel.acostamelgarejo@postgrad.manchester.ac.uk

- References:

  - Ambros, V. (2004). The functions of animal microRNAs. *Nature*, 431(7006):350355. Number: 7006 Publisher: Nature Publishing Group.
  - Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA 07, pages 10271035, New Orleans, Louisiana. Society for Industrial and Applied Mathematics.
  - Dongen, S., Dongen, v., Hazewinkel, M., and van Eijck, D. (2000). *Graph Clustering by Flow Simulation*. Universiteit Utrecht. Dongen, S., Dongen, v., Hazewinkel, M., and van Eijck, D. (2000). Graph Clustering by Flow Simulation. Universiteit Utrecht.
  - Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD96, pages 226231, Portland, Oregon. AAAI Press.
  - Fowlkes, E. B. and Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553569. Publisher: American Statistical Association, Taylor & Francis, Ltd.
  - Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl_1):D140D144. Publisher: Oxford Academic.
  - Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1):193218.
  - Slide 2 image: https://www.genengnews.com/magazine/december-1-2018-vol-38-no-21/microrna-profilers-cite-reconcilable-differences