MGKA: A genetic algorithm-based clustering technique for genomic data

Hung Nguyen, Sushil J. Louis, Tin Nguyen*

Computer Science and Engineering, University of Nevada, Reno

Contact: tinn@unr.edu, Website: https://bioinformatics.cse.unr.edu/

Backgrounds

- Advances in high-throughput technologies produces a huge amount of genomic data.
- High demand on finding precise disease subtypes from molecular measurement to reduce cases with over-diagnosis or under-diagnosis.
- Single-cell sequencing technology enables cell types discovery using gene expression data.

➔ substantial need to develop a clustering technique dedicated for genomic data.

- k-means, a broadly used and well-known clustering technique, was found to be efficient for clustering cancer datasets.

Problems

- k-means algorithm is sensitive to initial conditions and does not guarantee to produce global optimal clusters.
- The number of clusters must be given as an input parameter for the k-means clustering technique. Without any prior knowledge of the data, determining the appropriate number of clusters is considered a difficult task.

Challenges

- Finding the global optimal cluster for k-means algorithm and, at the same time, determining the appropriate number of cluster for genomic data without prior knowledge.

Our solution: Multi-objective Genetic algorithm- based K-means Algorithm (MGKA)

Use Multi-objective Genetic Algorithm to simultaneously optimize k-means solutions and find the appropriate number of clusters with Silhouette and Davies–Bouldin indices.

Multi-objective genetic algorithm-based k-means

- Real-number center-based encoding is used to present k-means solutions with dynamic number of clusters.

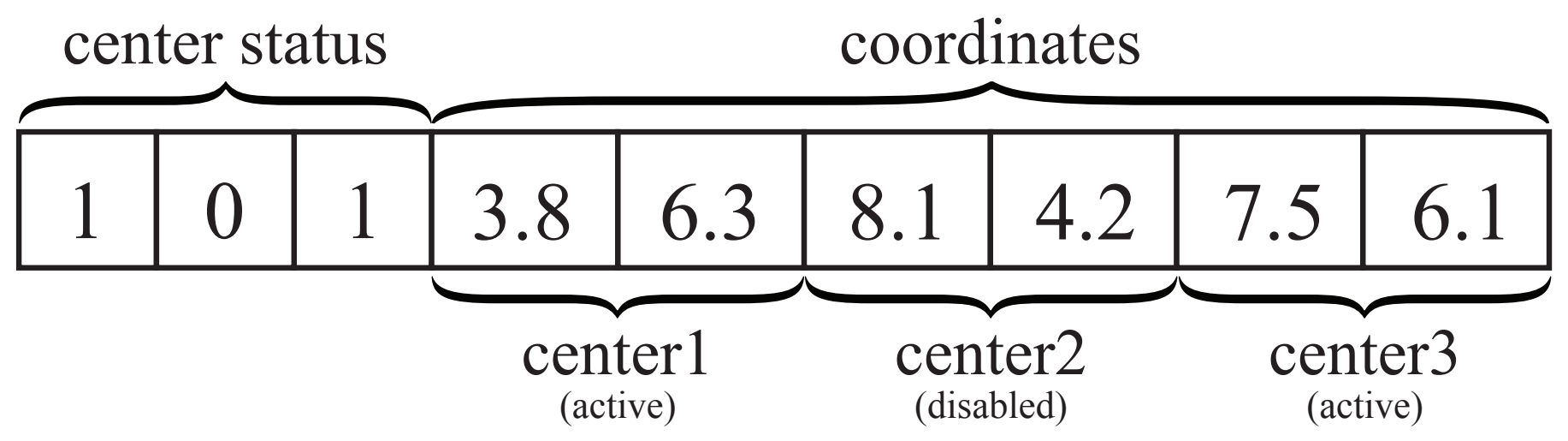

Fig. 1*: Chromosome encoding of a two-cluster solution for two-dimension data. By toggling the center status, the maximum number of clusters it can present is three.*

- Offspring are produced using simulated binary crossover [1]

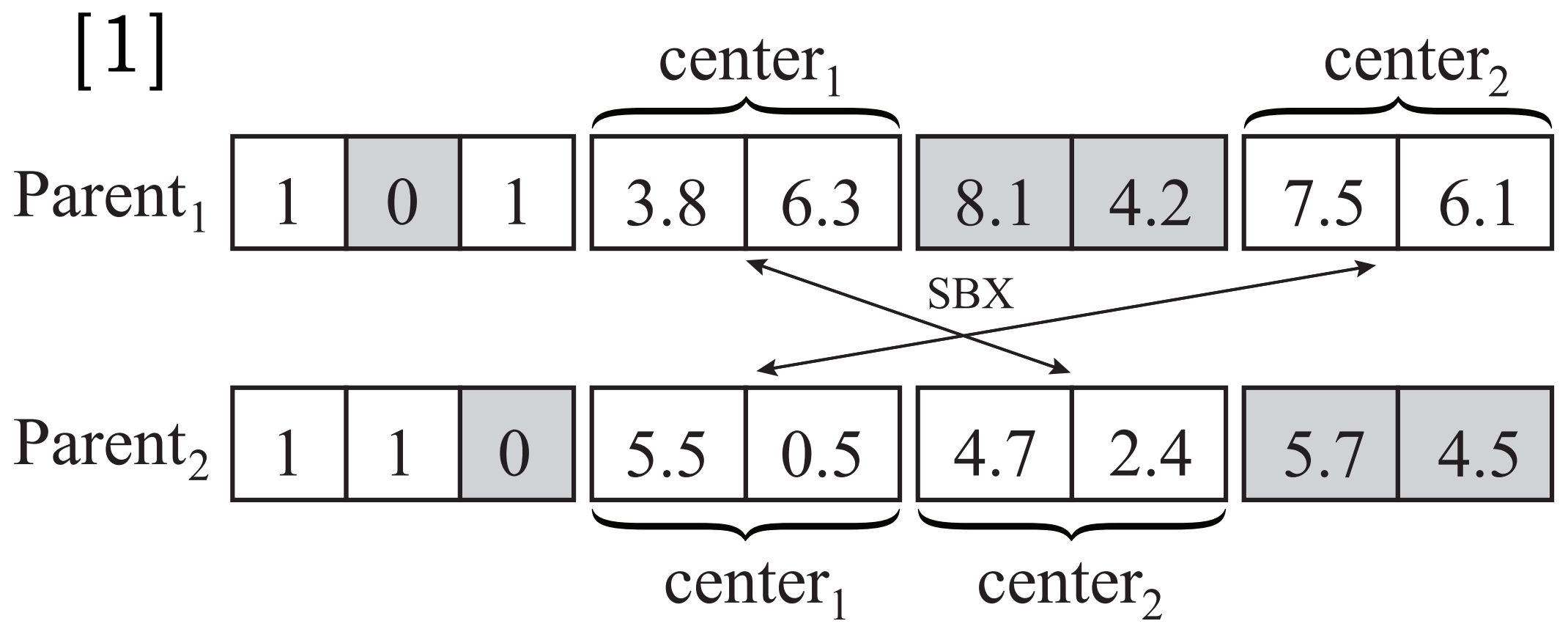

Fig. 2*: Simulated binary crossover procedure by two parents with the same number of clusters.*

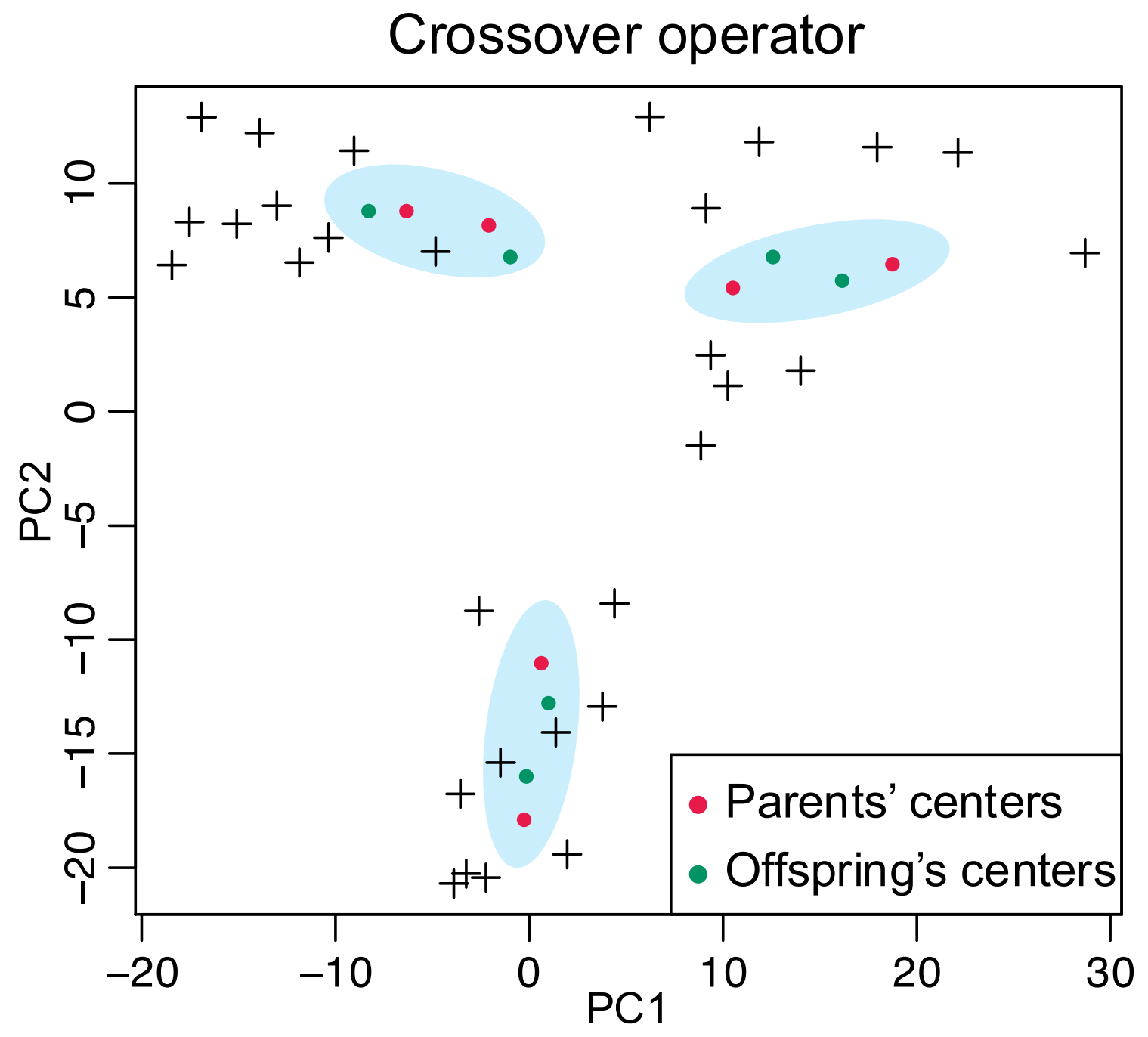

Fig. 3*: Offspring resulted from simulated binary crossover.*

- Within cluster sum of squares and two clustering indices are used to select the parents: i) Davies-Bouldin index: measures how well the clusters are separated, and ii) Silhouette index: measures how similar an object is to its own cluster compared to other clusters.
- NSGA-II [2] is used to optimize objectives in the selection operator.

Validation data

- We compare our method with the original k-means on simulated datasets with high number of clusters.
- Eight real disease datasets from Gene Expression Omnibus and Broad Institute with known subtypes are used to evaluate our method in comparison with other methods developed for disease subtyping including Similarity Network Fusion (SNF) [3], Consensus Clustering (CC) [4], and iClusterPlus [5].
- Four single-cell datasets with known cell types are used to evaluate our method in comparison with other methods developed for single-cell clustering including SC3 [6] and SEURAT [7].

Validation results

- We use Adjusted Rand Index (ARI) to measure the similarity between clustering results and the ground truth.

#k	#Samples	WithinSS		ARI	
		MGKA	k-means	MGKA	k-means
10	100	457.237	782.051	1	0.963
11	110	461.326	996.554	1	0.954
12	120	520.686	913.989	1	0.939
13	130	598.247	910.19	0.993	0.914
14	140	547.731	1136.477	1	0.931
15	150	630.188	1074.967	1	0.929

Table 1*: Performance of MGKA on simulated data*

Dataset	Samples	#Class	MGKA	CC	SNF	iCluster+
GSE10245	58	2	0.80	0.32	0.38	0.22
GSE19188	91	3	0.84	0.6	0.12	0.19
GSE43580	150	2	0.44	0.37	0.15	0.21
GSE15061	366	2	0.78	0.43	0.05	0.15
GSE14924	20	2	1.00	0.25	NA	0.73
Lung2001	237	4	0.54	0.11	0.28	0.11
AML2004	38	3	0.41	0.56	0.17	NA
Brain2002	42	5	0.15	0.46	0.13	0.32

Table 2*: Performance of MGKA on disease datasets*

Dataset	Samples	#Class	MGKA	SC3	SEURAT
Yan (GSE36552)	90	6	0.67	0.63	0.53
Goolam (E-MTAB-3321)	124	5	0.72	0.63	0.57
Deng (GSE45719)	268	6	0.60	0.55	0.51
Pollen (SRP041736)	301	11	0.88	0.93	0.70

Table 3*: Performance of MGKA on single-cell datasets*

References

1. Agrawal et al. (1995). Simulated binary crossover for continuous search space. Complex Systems, 9(2), 115–148.
2. Pratap et al. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation, 6(2), 182–197.
3. Wang et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 11(3), 333.
4. Monti et al. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. 52(1-2), 91–118.
5. Mo et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences, 110(11), 4245–4250.
6. Kiselev et al (2017). Sc3: Consensus clustering of single-cell rna-seq data. Nature methods, 14(5), 483.
7. Butler et al. (1028). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology, 36(5), 411.