

Usefulness of a Dendrogram Plot Technique in Clustering Homogeneous Subgroups of Multi-Treated HIV-1 Patients in Accordance with Their Mutational Pattern

Perez-Alvarez N^{1,2}, Libre JM¹, Clotet B^{1,3} for the Call conferences group.

1 Lluita contra la SIDA Foundation, Hospital Universitari Germans Trias i Pujol, Barcelona; 2 Universitat Politècnica de Catalunya; 3 IrsiCaixa Foundation, Barcelona. Spain



AIM

To investigate the utility of a **multivariate clustering statistical technique** in the characterisation of **homogeneous subgroups of multi-experienced patients** with virological failure. This characterisation could identify homogeneous subgroups susceptible of further resistance analysis.

METHODS

The pattern of resistance-associated mutations (RAMs) of **195 genotypes from heavily multi-treated patients with virological failure** was studied by means of a dendrogram plot and **k-means cluster analysis**. The pharmacological history was evaluated for each of the mutation-clustered subgroups.

The set of patients studied had a median (IQR) CD4 cell count of 282 (252) cells/mm³, HIV-1 viral load of 4.2 (1.4) log₁₀copies/mL, and 5 (2) lines of previous HAART regimens.

K-means method

The k-means method will produce exactly k (number of groups) different clusters of greatest possible distinction. It should be mentioned that the best number of clusters k leading to the greatest separation (distance) is not known a priori and must be computed from the data. Here is step by step the k means clustering algorithm:

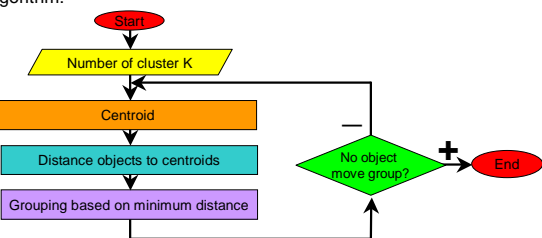


Figure 1. K-means clustering algorithm.

K=? There is no general theoretical solution to find the optimal number of clusters for any given data set. A simple approach is to **compare the results of multiple runs with different k classes** and choose the best one according to a given criteria (for instance the Schwarz Information Criteria). However, we need to be careful when choosing the number of groups because the higher increase in the number of clusters (k) the lower the information retrieved.

The optimal k to choose for further working should be the first one after the quickly decrease of the cost function (in this case, the Schwarz Information Criteria) and before the flat set point reached when number of clusters increases; but it must be coherent with clinical judgement.

RESULTS

Dendrogram for the mutational pattern

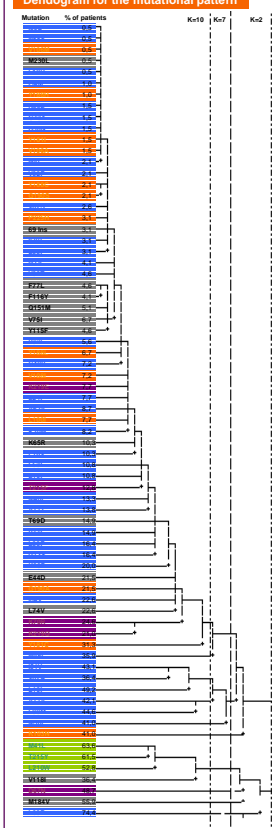


Figure 2. At the dendrogram for the mutational pattern, the protease mutations, TAM1, TAM2 reverse transcriptase mutations and mutations associated to NNRTIs are identified by colours, and the percentage of patients harbouring every mutations is reported.

Mutations grouped in every arm depict the closest similarities between them (homogeneity). Distant mutations (the not connected directly by any arm) are highly heterogeneous.

K is the number of clusters (homogeneous groups).

Top 15 associated mutations identified and the pharmacological history of each cluster

Cluster 1				Cluster 2			
Mutation	Centroid	Drug experience	%	Mutation	Centroid	Drug experience	%
K10N	0.50	3TC	100.0	L10I	0.68	EFV	100.0
D67N	0.90	dSd	95.2	M18V	0.55	AZT	91.7
L33P	0.81	d4T	95.2	K10N	0.42	dSd	91.7
V32A	0.78	AZT	90.0	D67N	0.28	d4T	91.7
A71V	0.71			M30I	0.22	3TC	91.7
L50M	0.71			Y181C	0.19	TPV	83.3
K219Q	0.67			M46I	0.14	LPV	83.3
L10I	0.67			A71V	0.13	IDV	83.3
M46I	0.62			L10I	0.10		
M30I	0.62			K20R	0.09		
T215V	0.57			IS4V	0.08		
IS4V	0.57	ATV	9.5	K66R	0.06	TPV	33.3
Q190A	0.48	T20	9.5	V32A	0.04	T20	33.3
M18V	0.38	APV	4.8	M46I	0.01	95C	16.7
L74V	0.38	TPV	4.8	Q151M	0.00	ATV	16.7

Cluster 3				Cluster 4			
Mutation	Centroid	Drug experience	%	Mutation	Centroid	Drug experience	%
L33P	0.88	3TC	95.7	IS4V	0.94	3TC	100.0
T215V	0.81	d4T	84.1	A71V	0.84	LPV	100.0
M46I	0.59	AZT	78.7	L33P	0.84	dSd	93.8
M18V	0.55	dSd	78.7	V32A	0.81	d4T	93.8
L210W	0.43			M18V	0.81	AZT	87.5
K10N	0.42			M46I	0.75	SQV	87.5
D67N	0.28			F55L	0.63		
M30I	0.22			L10I	0.69		
V118I	0.20			Q178S	0.50	APV	18.8
Y181C	0.19	95C	14.5	M46I	0.50	ATV	18.8
V77I	0.17	APV	11.6	L210W	0.50	TPV	18.8
L50M	0.14	ATV	7.2	M46I	0.44	TPV	12.5
L74V	0.14	T20	4.3	V118I	0.44	T20	12.5
M46I	0.14	TPV	2.9				

Cluster 5				Cluster 6			
Mutation	Centroid	Drug experience	%	Mutation	Centroid	Drug experience	%
M46I	1.00	dSd	97.2	K219Q	0.92	d4T	92.3
T215V	0.84	d4T	97.2	L33P	0.86	3TC	92.3
L210W	0.89	3TC	97.2	K76R	0.77	AZT	84.6
L50M	0.86	LPV	86.1	Y181C	0.69	dSd	84.6
L33P	0.81	IDV	86.1	M18V	0.62	NEV	84.6
L10I	0.75	AZT	83.3	K10N	0.54	IDV	84.6
M18V	0.72			L10I	0.54		
IS4V	0.69			D67N	0.48		
IS4V	0.69			M46I	0.46	95C	30.8
D67N	0.64			L50M	0.38	LPV	30.8
L74V	0.64			K55E	0.38	TPV	30.8
Y181C	0.58	95C	33.3	F77L	0.38	T20	30.8
L33P	0.58	T20	25.0	V72F	0.38	TPV	15.4
A71V	0.58	ATV	22.2	IS4V	0.38	ATV	7.7
E44D	0.56	TPV	11.1				

Cluster 7			
Mutation	Centroid	Drug experience	%
L210W	1.00	dSd	100.0
M46I	0.96	d4T	100.0
T215V	0.93	AZT	98.4
V32A	0.79	3TC	92.9
L33P	0.79	RTV	89.3
V118I	0.75	LPV	85.7
M30I	0.68	TDF	82.1
D67N	0.68	SQV	82.1
L10I	0.68		
A71V	0.67		
IS4V	0.57	TPV	32.1
K20R	0.54	T20	28.6
L33P	0.46	APV	25.0
E44D	0.36	ATV	25.0

Table 1. At the sub-tables, for the homogeneous groups defined by the clustering analysis, the largest 15th coordinates of the centroid are reported. Mutations holding a larger coordinate are representative of the cluster.

These combinations of mutations are the most representative for every homogeneous group. The drugs with largest and lowest representation in each cluster are listed, also the percentage of patients who had been take each drug is reported.

Pharmacological history and HIV-Clinical Markers for each cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
AZT	90.0	91.7	79.7	87.5	83.3	84.6	95.4
d4I	95.2	91.7	79.7	93.8	97.2	84.6	100.0
d4C	33.3	16.7	14.5	37.5	33.3	30.8	50.0
d4T	95.2	91.7	84.1	93.8	97.2	92.3	100.0
3TC	100.0	91.7	95.7	100.0	97.2	92.3	92.9
ABV	57.1	58.3	46.4	62.5	58.3	53.8	64.9
TDF	71.4	83.3	40.6	81.3	75.0	76.9	82.1
NEV	38.1	41.7	63.8	75.0	75.0	84.6	67.9
EFV	66.7	100.0	60.9	68.8	61.1	53.8	71.4
LPV	66.7	83.3	36.2	100.0	86.1	30.8	85.7
RTV	78.2	66.7	50.7	81.3	72.2	69.2	89.3
SQV	57.1	58.3	36.2	87.5	69.4	38.5	82.1
IDV	71.4	83.3	62.3	75.0	86.1	84.6	78.6
NFV	4.8	41.7	11.6	18.8	41.7	38.5	25.0
APV	4.8	41.7	11.6	18.8	41.7	38.5	25.0
ATV	9.5	16.7	7.2	18.8	22.2	7.7	25.0
TPV	4.8	33.3	2.9	18.8	11.1	15.4	32.1
T20	9.5	33.3	4.3	12.5	25.0	30.8	28.6
CD4+(cells/mm ³)	270 (151:450)	282 (155:415)	167 (124:294)	334 (218:469)	346(184:601)	271 (198:408)	265 (137:408)
VL (log ₁₀ copies/mL)	4.3 (3.7:4.9)	4.1 (3.2:4.8)	4.0 (3.1:5.3)	4.0 (3.2:4.5)	3.6 (3.2:4.3)	4.4 (3.7:4.8)	4.5 (3.9:5.1)
Time infected (years)	13 (9:15)	12 (6.5:15)	13 (10:16:25)	13 (10:16)	12 (9.5:14)	13 (12:15)	14 (11:16)
Time on ART time (years)	13 (9:11)	8 (6:10)	9 (7:10:25)	9 (8:10:75)	9 (7.5:10)	8 (8:10)	9 (8:11)

Table 2. The pharmacological experience is described by the percentage of patients belonging to each cluster who had been taken each drug. Median (Percentile 25%; Percentile 75%) of the clinical markers is also reported. Green colour indicates the cluster having the 2 larger proportion of the drug and blue indicates the cluster having the 2 smaller proportion of the drug.

CONCLUSION

The **dendrogram** plot and **k-means cluster** analysis effectively identify homogeneous subgroups of heterogeneous multi-treated HIV-1 patients with treatment failure according to the RAMs harboured.

This statistical technique is widely used to find homogeneous patterns in multi-dimensional heterogeneous datasets. Identifying groups with clustered mutations that emerge together may be useful to undertake further treatment or resistance analysis in this scenario.

CALL CONFERENCES GROUP

Arao P, Cervantes M, Clotet B, del Pozo MA, Domingo P, Galindo J, Gutierrez M, Iribarren J, Libre JM, Marino A, Miralles C, Moreno S, Ocampo A, Puig T and Schapiro J.

ACKNOWLEDGEMENTS

To Cecilia Alcaraz, for her efficiency in coordinating the Call Conference group.