https://creationismeweersproken.blogspot.com/2023/01/de-rode-panda-en-cserhati-8-clustering.html

THE RED PANDA AND CSERHATI (8): CLUSTERING

Cserhati employs two techniques for red panda and giant panda placement based on the Whole Genome K-mer Signatures: a phylogenetic tree and clustering.

Cserhati does not put much emphasis on the phylogenetic tree based on his WGKS data of 28 species, but gives ample attention to clustering. In his analysis he creates a correlation matrix, the pairwise correlations of the octamer signatures of the species. He displays this matrix in a 'heat map' in which the size of the correlation is shown on a light-dark scale. Based on this matrix, Cserhati searches for clusters.

> *The last step (in the analysis) involves visualizing the PCC in a heatmap and using clustering algorithms to detect monophyletic groups.*

"Using clustering algorithms to detect monophyletic groups".

Really? You're serious?

Is it possible to find or define groups of common descent, monophyletic groups, from clustering?

1 Clustering describes similarity, not phylogeny

In clustering, we have a large number of independent individuals (persons, schools, cars, pieces of music, countries) each with a number of characteristics. In clustering we look for 'who looks like whom', whether groups can be defined so that each individual within the group resembles every other individual in that group more than an individual in the other group. Such a group is called a cluster. A cluster indicates similarity. Lineage is not an issue in clustering. Nor is monophyly: that is common descent, not clustering.

2 A phylogeny yields clusters

If we have a phylogeny, and choose species from it: eg 5 monkeys, 5 rodents, 5 cattle, 5 bats, and run a clustering program, we are guaranteed (here) 4 clusters, each with 5 species. We introduce existing groups, and when existing

known groups are put into a clustering program, those known groups will be found.

In the phylogeny of the carnivore order Carnivora, we have the cat family Felidae, bear family Ursidae, and the superfamily Musteloidea. When Cserhati submits these 28 species of carnivores belonging to these monophyletic groups to clustering, he finds three clusters: cats, bears and Musteloidea. Cserhati put species from three monophyletic groups in a cluster analysis, and of course found those monophyletic groups.

3 A monophyletic group gives a monophyletic cluster

Not the other way round. A group of species is not monophyletic because they cluster together, they cluster because they are monophyletic.

4 Monophyly cannot be concluded based on clustering, .

'Cluster' and 'monophyletic group' will often coincide when biological species are being clustered, but when clustering some seal species with *Poiana leightoni*, *Poiana richardsonii* (Africa linsangs), *Prionodon linsang* and *Prionodon pardicolor* (Asian linsangs) we (presumably) get two clusters, one with the seals and one with the linsangs. This while the Asian linsangs and the African linsangs do not belong to the same family or superfamily. Such a linsang cluster is heterogeneous. Nothing in clustering per se tells us that a cluster would be monophyletic.

5 Clustering and Input

The number of clusters depends on the scope of the input. Among the 28 species with WGKS data from Cserhati, there is a cluster of cats, a cluster of bears and a cluster of Musteloidea. There are 12 species in the Musteloidea cluster and this cluster does not split any further. In Cserhati's analysis on mtDNA there are 37 species of Musteloidea, and the Musteloidea split into 4 clusters: viz the four monophyletic families Mephitidae, Ailuridae, Procyoniidae and Mustelidae - giving clusters because they are a monophyletic family. The 10 species of Mustelidae do not split further into clusters, but just entering many Mustelidae species can lead to different clusters.

6 A phylogeny is hierarchical, clustering is not

Clustering cannot discern a hierarchical structure in the data. The hierarchy in the classification of the living creatures appears only on repeating the clustering input of different scopes (see example under point 5). This means that clustering of biological groups does not give a good representation of the hierarchical structure of the animal world.

7 Clusters have nothing to do with relatedness

Clusters only give the optimal split of the data, nothing about relatedness, neither within nor between clusters. Relatedness follows from the phylogeny.

For examples, see points 4 and 5. The family Mustelidae is monophyletic and the species are related, but if you put otters, weasels and martens in a clustering program you will find several clusters. The families of the monophyletic superfamily Musteloidea are related, but in a clustering program they emerge as a cluster given sufficient scope of input. If only all cats species are put in a cluster program, you get a cluster of 'big cats' and a cluster of 'little cats'. The 'big cats' and the 'little cats' are related, even though they end up in different clusters

8 Clustering is a statistical trick, not a biological classification

A phylogeny is biology, clustering statistics.

Summarizing, "*using clustering algorithms to detect monophyletic groups*" shows no insight into clustering or phylogeny or biology. Clustering in the same cluster cannot be used to conclude to monophyly and relatedness. Clustering in two clusters cannot be used to conclude the species in the separate clusters are not related.

****

Cserhati, M., 2021, A tail of two pandas – whole genome k-mer signature analysis of the red panda (Ailurus fulgens) and the Giant panda (Ailuropoda melanoleuca), BMC Genomics 22: 228

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07531-3