Case Study on Clustering

Xiaocheng Li
Imperial College Business School

MNIST data K-means clustering

MNIST data:

http://yann.lecun.com/exdb/mnist/

Blog:

https://medium.com/datadriveninvestor/k-means-clustering-for-imagery-analysis-56c9976f16b6

Genomic Applications

In computational linguistics and computer science, edit distance is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other.

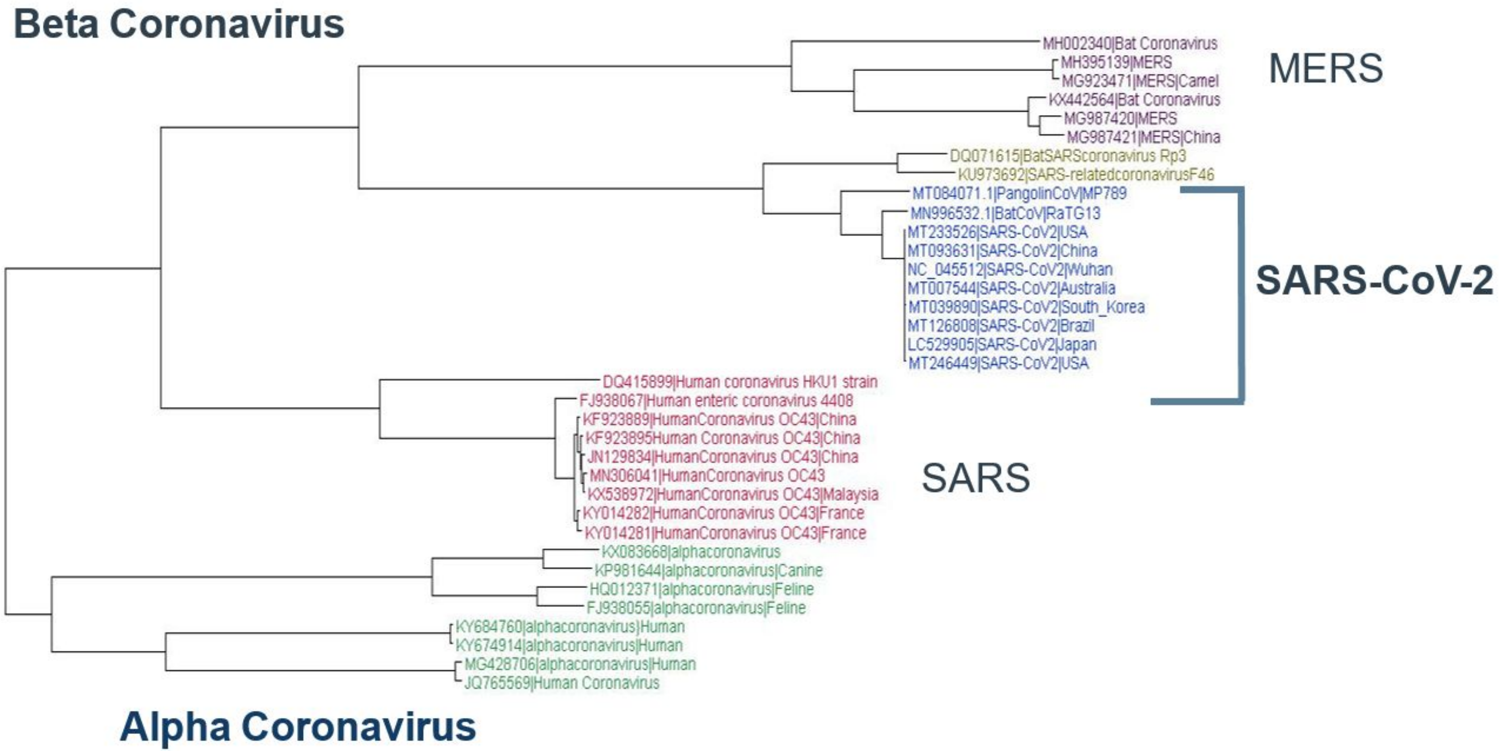
The operations including insertion, deletion and substitution

Example:

d(ACGTC, ACTC) = 1

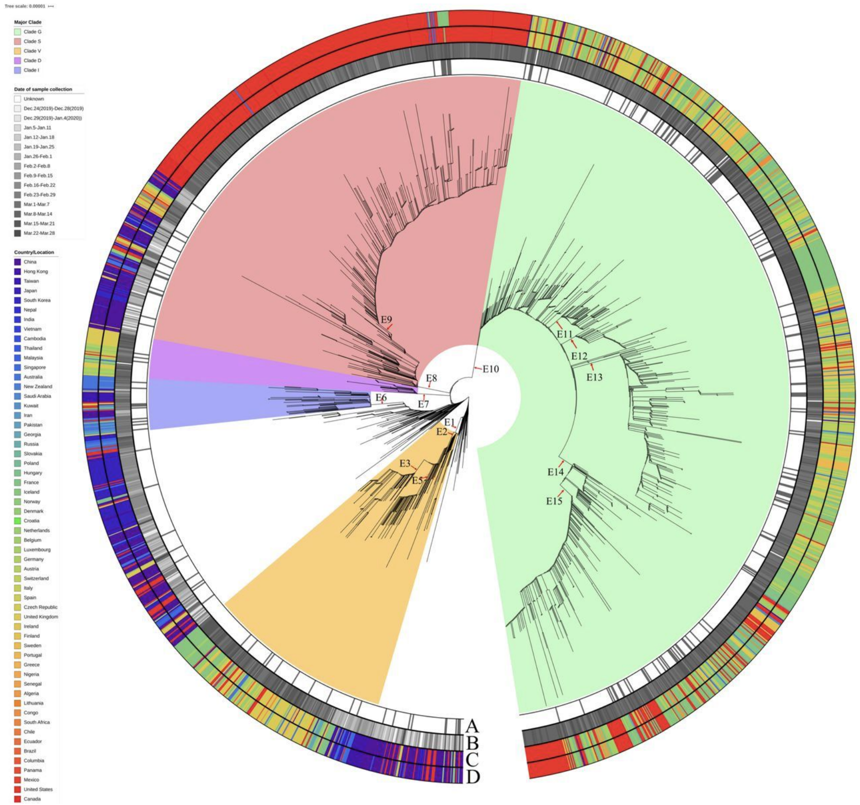
d(ACGTC, CGATC) = 2

Hierarchical Clustering Analysis of Covid-19 (I)



Beta Coronavirus
MH002340|Bat Coronavirus
MH395139|MERS
MG923471|MERS|Camel
KX442564|Bat Coronavirus
MG987420|MERS
MG987421|MERS|China
MERS
DQ071615|BatSARScoronavirus Rp3
KU973692|SARS-relatedcoronavirusF46
MT084071.1|PangolinCoV|MP789
MN996532.1|BatCoV|RaTG13
MT233526|SARS-CoV2|USA
MT093631|SARS-CoV2|China
NC_045512|SARS-CoV2|Wuhan
MT007544|SARS-CoV2|Australia
MT039890|SARS-CoV2|South_Korea
MT126808|SARS-CoV2|Brazil
LC529905|SARS-CoV2|Japan
MT246449|SARS-CoV2|USA
SARS-CoV-2
DQ415899|Human coronavirus HKU1 strain
FJ938067|Human enteric coronavirus 4408
KF923889|HumanCoronavirus OC43|China
KF923895Human Coronavirus OC43|China
JN129834|HumanCoronavirus OC43|China
MN306041|HumanCoronavirus OC43
KX538972|HumanCoronavirus OC43|Malaysia
KY014282|HumanCoronavirus OC43|France
KY014281|HumanCoronavirus OC43|France
SARS
KX083668|alphacoronavirus
KP981644|alphacoronavirus|Canine
HQ012371|alphacoronavirus|Feline
FJ938055|alphacoronavirus|Feline
KY684760|alphacoronavirus|Human
KY674914|alphacoronavirus|Human
MG428706|alphacoronavirus|Human
JQ765569|Human Coronavirus
Alpha Coronavirus

Hierarchical Clustering Analysis of Covid-19 (II)



Bi-Clustering

The clustering algorithm helps us understanding the relation between samples

That is, we treat the feature of each sample as a vector, and perform hierarchical clustering for the samples

Also, we can do the same thing for the features, and the hierarchical clustering helps us understanding the relation between the features

Of course, we can do both at the same time

https://docs.google.com/document/d/1XIAVIO3kHRIJv0-0kadrXV9UTU8EzUMt7NxUS-E39yI/edit

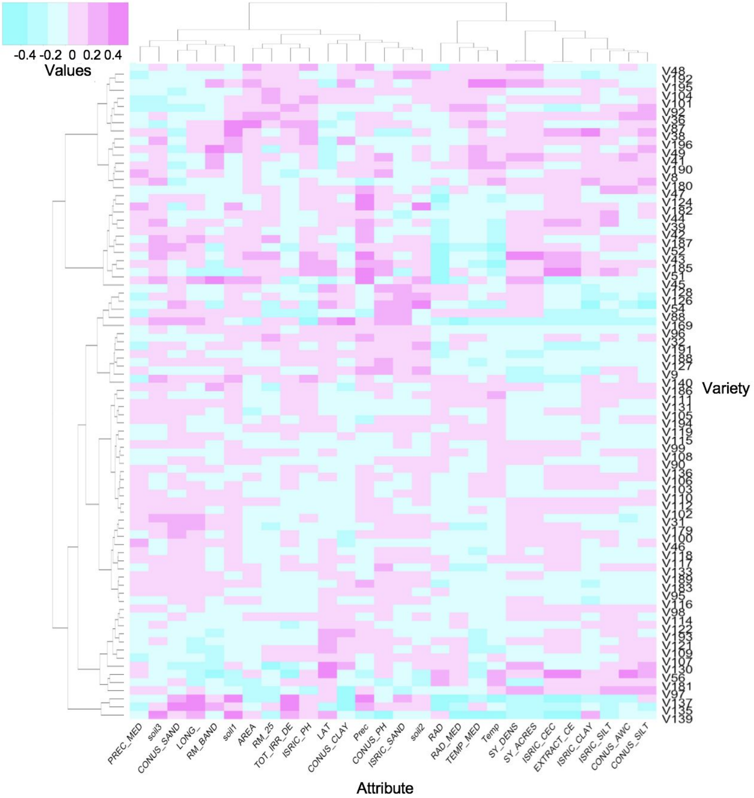
Example of Biclustering

https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0119448.g005

Each row: A gene

Each column: A condition

Agriculture application



A dataset of crop yields across hundreds of farmlands in U.S.

100 different varieties/typies of the crop

Each farmland is monitored by 20 features

Left panel: A matrix of correlations, each correlation is

Clustering and Predictions

Predicting tomorrow’s demand for a certain product

Important features:

- Demand of the today
- Average demand of the past seven days
- Demand of the today - 7
- Demand of the today - 14
- Price of the product
- Price of the complementary/substitute
- ...