CODE

done

- working functions
  - readcods
  - retrieve_list
  - preprocess yun
- to test
  - load and save
  - put back the pandas dataframe
  - homologize from matrix

coding

- 6 download the data files
- 7 show the clusterisation in bokeh.

todo

- find tests and visualisations to perform on the clustering of the homology matrix
- add a verification that the function are not called in a wrong order

later

- change the pipeline according to the biopython package
- create a CDC for the pipeline
- I should look at species that are not represented in the dataset.
- I can try to retrieve the raw sequences which names are given in the second row of the dataset and match them to the database (online sql queries OR download it and do direct sql queries ensembl.org)
- plot the number of species found per genes.

INFO

ideas

- see if there is a similarity in the cluster position of each of their genes, compare it and find relationships
- compare the similarities between the species for each types of related genes between the species
- find a relationship between
- plot entropy location values similarity between genes of one particular individual
- look for philogenetic similarity using ncbi taxonomy information/website or the information for

Tobias
- To find Metadata
  - look at the text from ensembl and do a feature extraction using NLP
  - look for websites

other things to do

- the species are ordered according to philogenetic/taxonomic similarities send an email to tobias for the name of the comparer species. ask for the philogenetic list amongs the 461 species the species present here are all the same ones.
- find how to replace Nans as they convey info and couldnot be replaced by 0 or 1 (maybe 0.5 ?) `http://scikit-learn.org/stable/modules/preprocessing.html`
- look at the equations on the thesis.
- WE can look for classification methods such as :

CLUSTER

- K-means
- Gaussian Mixture (clustering) http://scikit-learn.org/stable/modules/mixture.html#gaussian-mixture
- Affinity propagation (cl)don't forget performance testing of clustering algo http://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
- DB-scan

CLUSTERER-CLASSIFIER

- self organizing maps > 1000 neurons NOVELTY/OUTLIER DETECTOR http://scikit-learn.org/stable/modules/outlier_detection.html

FEATURE EXTRACTOR

- Var autoencoder
- graphical models http://scikit-learn.org/stable/modules/neural_networks_unsupervised.html#graphical-model-and-parametrization

DIM REDUCTION

- PCA - ICA
- tsn-e

OTHER

- NMF (given this gene values, to which species it belongs| given this gene cluster, to which species it belongs | given this species what could be the value if it had this homology) http://scikit-learn.org/stable/modules/decomposition.html#non-negative-matrix-factorization-nmf

-or-nnmf

Pipeline of the matlab code

so you have one species and you look for homologies to a subset of 5818 genes of this species. th main = homologyvalue() then you gethomoinformation() to get the information about the homology to 461 species then you look at preinstalled genotype of 461 species and get this particular gene

- read the paper
- cancerous cells should have a particular kind of codon usage bias (more than regular cells)
- Entropy location : you compute the entropy value for each possible configuration, it makes a distribution to wich you compare the measured entropy value. with a integration.
- there is a whole genome part as well KL method : for the comparison of the whole genome per species (only exons !) we sum up all the distributions and normalize them (by the length of the gene) and we compute the KL divergence to the distribution made by the measured distribution made by the plotting of all the exons Expected entropy: (see notebook)
- focus first on the classification
- be really carefull about what you are goin to present on your thesis.

communication management

Dominique Chu

> Really do the same plot with the 5000 species to look for structure Do it with entropyValue and Entropy*Length as well

Alex Freitas

- density based clustering (and ensemble)
- look for way to assess clustering quality (inter intra cluster)
- regularize by the amount of clusters (pareto)

> the gene clustering (of i-one gene) can be seen as a mini phylogenetic tree once I have the meta data how much can each features or groups of features can explain the clustering that I found.

question

- how are you finding your homologies I feel as though they are not the same for every species. How can I compare them in this way ?

> you should not look at the homologies like that. the genes can actually be very different

and the way they find homologies is by comparing how it behaves or if they know it is an homologies or the RNA or DNA list...

- should I normalize the data using the lenghth to compare them ?

  no

- how do I read your csv ?

  it is ordered as a tree from similarity alignment matching (look for similarity alignment phylogenetic matching algorithm) BUT for Yun's it is according to the names (family) taxonomical data + similarity alignement matching (just getting it from itol.embl.de)

- where can I get the philo tree that you have ?

  to recreate it from the CSV file you need to send it to the itol.embl.de website

- unicel or not ?

  not

- pathogene or not (plant/animal) ?

  some yes

To say

problem on what I am going to write and do as a master thesis python package with full pipeline (versatile) how much can I say about yun's work (she was stressed)

homologues not related to theur functions but their common ancestors the way

require from Tobias

- phylogenetic tree
- temperature
- replication speed
- frequency of use of the gene
- nocive species
- gene size

- type of species
- more ...

work pipeline

Phase (1) – Business understanding Understanding the business objectives and requirements, and converting them into a data mining problem

Phase (2) – Data understanding Understand the data to be mined, consider data quality issues

Phase (3) – Data preparation Data cleaning, attribute and record selection, data transformation, etc.; dependent on the data mining algorithm(s) to be used in (4)

Phase (4) – Modelling / data mining phase Choose one or more data mining algorithms to be applied to the data, adjust its(their) parameters, build a model of the data Includes an initial objective, data-driven validation

Phase (5) – Model evaluation Interpret and validate the models from a business perspective

Phase (6) – Deployment of the model Deploy the data mining results in the form of a plan, monitoring and maintaining the plan deployment