

# PyCUB: Notebook

## 2017-2018

Jérémie kalfon  
jkobject@gmail.com

This document contains a slightly edited -although uncorrected- copy of my logbook and some more images of my work on PyCUB.

## December 2017

## Writting

End november

I went to see Dominique

I have been asked to create a clustering and classification algorithm of codon frequency usage (entropy) for a PhD student that needs to use ML on her dataset and do not have strong knowledge in CS.

Thanks to my previous knowledge in CS, ML and Biosciences, I have been able to understand the problematics and come up with some ideas.

Looking at the problematic and assessing the project.

I have been given a warning about a time limit on the project. Yun needs to finish her PhD ASAP and we have to publish in march.

begin december

I went to see Dominique

I realised I needed more information about the problematics and what was the project main ideas, output and realisations.

I researched the literature

mid december

However to do that I needed to have many many discussions with YUN and Tobias which have the knowledge I and Dominique don't I have been able to understand better some key facts about the data I have been given. I realized that understanding the data was really important

I went to see tobias

---

I went to see Yun  
I produced a first draft code during the holidays

## First Draft

### *First mindmap*

- classification
  - Using K-means
- classifier --NN
  - to infer with metadata
- metadata
  - phylogenetic tree
  - temperature
  - replication speed
  - frequency of use of the gene
  - nocive species
  - gene size
  - type of species
- dimensionality reduction ?
  - create a higher dimensionality dataset (using FOC instead of entropy)
  - PCA

### *Technical information*

#### **Different types of codon usage bias measurements :**

Codon Adaptation Index N → since it uses an idea on why there is a frequency bias and it uses other codons.

(FOC) frequency of optimal codons : to implement ( Higher Dim)

Relative Codon Adaptation :

effective number of codons :

Shannon entropy H

#### **Why is there CUB:**

- Cause :
  - Mutational bias versus selection
  - GC
  - Temperature
  - replication speed
  - Length of the gene

- 
- Interaction with the tRNA pool (thus with its expression and the one of all the proteins of the cell)
  - Consequences
    - change transcription efficacy by changing RNA structure
    - Modulation on the speed of translation. Which temporally -regulates behavior and which could change the proper folding of the protein

## QUESTIONS

What's the topic area?

the topic area certainly are Machine Learning, Biosciences, Genomics, Computer Science, Data Science.

What's the context?

The context is that this project is the fruit of a work done by Yun Deng. The CUB is not very much researched and its explanation are not known exactly.

I am very much interested by genomics and ML and Dominique Chu saw it. That is why he proposed this project. BioPython is a big Bioscience and Genomic Package for python that could be used

What are the constraints?

There is obviously a constraint of time. Furthermore, Yun Deng needs to finish her PhD ASAP. In addition, there is for now a constraint of data and computing power. we don't really what the computation will be and how much data we will have. Moreover there is a lack of metadata on the species that could help the algorithm infer the reason of the codon usage bias.

Do computation on the data :

(checking for data structure) -- how to mix them? confidence interval on effectiveness?

→ k-means

→ DBLearning

<http://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>

→ PCA ( need IPCA ? see size of dataset)

---

## January 2018

### Writting

mid january

I went to see Tobias

I went to see Yun

I changed the strategy according to my meetings : need to change the code entirely, need to do the entire pipeline, need to explain things more

End january

I went back to see yun and got knowledge of all the process. Reassure her about my intention to help here and tried to find the best possible course of action throughtout the semester

I skyped Dominique, he was the one to tell me to be quick in my pipeline and not to focus on detail (but it doesn't work like that with code) He said that I don't need to produce good code and that I don't need to write a logbook. And that I will only be graded on my thesis and my understanding of the research → contradiction with my view about how to do science and how to produce reproducible science.

→ very important in datascience (what we are doing here basically)

I try to find out what is the best pipeline and to get my head around the code conceptually. → lead to more questions and more focus on different problem that will arise.

## February 2018

### Writting

MID FEBRUARY

A lot of work but I manage to see tobias and Yun one more time which solved most of my final questions. However I am now left with much indecision about what to do with this project as it seems I am only left alone with no real goal except "Use ML and Help as you can we don't understand this stuff".

I am now planning different strategies. It helped me to go see Pr. Freitas for the help and to ask him questions more related about ML and how I could do this stuff.

END FRBRUARY

---

Starting to code according to my analysis and my understanding of the problem.  
First i am focusing on the most important part while keeping a structure that is made for the full python package should work by the end of february

Mid March

Broke my computer at the beginning of March and thus it is impossible to work on developing the project.

Focusing on reading and getting new ideas as the pipeline is clear but the possible outcomes are sparse.

Taking a lot of times to find relevant litterature and links for more data related topics

---

## Links

[https://en.wikipedia.org/wiki/Codon\\_usage\\_bias](https://en.wikipedia.org/wiki/Codon_usage_bias)

Main codon usage ressource

Cours UE génomique L3 ENS LYON

<http://genomes.urv.cat/CAIcal/>

Server to perform CAI

<http://gcuu.schoedl.de/>

Graphical codon usage analyser

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2174420/>

ACUA paper ( automated codon usage app )

<http://genomes.urv.cat/HEG-DB/>

Database of the highly expressed genes (for metadata)

<http://www.cbrc.kaust.edu.sa/CAT/#executables>

Codon usage analysis shell software (can run command line from python)

<http://codonw.sourceforge.net/>

Toolkit for codon usage in the command line containing a multitude of analysis possibility (CAI, RCA, GCcontent ...)

<https://www.ncbi.nlm.nih.gov/guide/taxonomy/>

Taxonomical database

---

<http://www.kentinformatics.com/products.html>

Blat program

<https://www.ensembl.org/Multi/Tools/Blast?db=core>

Blasted genes database

[https://en.wikipedia.org/wiki/List\\_of\\_sequence\\_alignment\\_software](https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

List of sequence alignment softs

<https://stackoverflow.com/questions/2352181/how-to-use-a-dot-to-access-members-of-dictionary>

Using . for dictionary

<http://epubs.siam.org/doi/10.1137/S089548010240415X>

A way to approximate max clique

[http://ipython.org/ipython-doc/stable/parallel/parallel\\_intro.html#examples](http://ipython.org/ipython-doc/stable/parallel/parallel_intro.html#examples)

Parallel computing in python (to dramatically increase the pipeline of this project)

<https://www.ncbi.nlm.nih.gov/pubmed/?term=unification+of+protein+abundance+datasets>

Paper on protein abundance with information about one species

<http://ensemblgenomes.org/info/data>

Information on the data provided by ensembl

<https://www.ensembl.org/info/website/upload/gff.html>

<http://gmod.org/wiki/GFF3>

Gff3 file format explanation

<https://academic.oup.com/nar/article/32/17/5036/1333956>

An important paper on CUB.

[https://en.wikipedia.org/wiki/Transfer\\_RNA](https://en.wikipedia.org/wiki/Transfer_RNA)

Important to know for the paper

<http://ieeexplore.ieee.org/document/6112394/>

Interesting paper on how to do a codon2vect basically in an effective way

<https://link.springer.com/article/10.1007%2F978-1-4939-9288-2>

Information on conservation in protein evolution

---

## *Interesting genetical packages in python*

<http://daler.github.io/gffutils/#example-file>

Reading gff3 files

<https://pypi.python.org/pypi/pysplicer>

A frequency-bias codon optimisation script with early NGG avoidance and optional IUPAC match avoidance.

<https://pypi.python.org/pypi/genetic-collections/0.1.7>

A Python library for connecting genetic records with specimen data.

<https://pypi.python.org/pypi/Geeneus/0.1.9>

Simple API for NCBI database access

<https://pypi.python.org/pypi/gnomic/1.0.1>

A grammar for describing microbial genotypes and phenotypes

<https://pypi.python.org/pypi/graftm/0.10.1>

GraftM is a pipeline used for identifying and classifying marker gene reads from metagenomic datasets

<https://pypi.python.org/pypi/fastac/0.2>

Compiler for FASTA files and a FASTA-based DNA scripting language.

<https://pypi.python.org/pypi/biothings/0.1.0>

a toolkit for building high-performance data APIs in biology

<https://pypi.python.org/pypi/alfpy/1.0.5>

Alignment-free package to compare DNA/RNA/protein sequences (bioinformatics)

<https://pypi.python.org/pypi/rnacounter/1.5.2>

Rnacounter estimates abundances of genes and their different transcripts from read alignments. Exons and introns can also be quantified.

<https://pypi.python.org/pypi/PyPhyloGenomics/0.3.12>

Tools to work in phylogenomics, from NSG group <http://nymphalidae.utu.fi>

A package to work on Phylogenomics.

<https://pypi.python.org/pypi/phylogenetics/0.4.1>

Python API for managing a phylogenetics project

<https://pypi.python.org/pypi/phylotoast/1.3.0>

Tools for phylogenetic data analysis including visualization and cluster-computing support.

<https://pypi.python.org/pypi/popsc/1.0.1>

PopSc is a toolkit for computing 45 basic statistics in molecular population genetics, mainly including (i) genetic diversity of DNA sequences, (ii) statistical tests for neutral evolution, and (iii) measures of genetic differentiation among populations.

---

<https://pypi.python.org/pypi/drawm/0.0.2>

A toolkit for creating publication-quality images of phylogenetic trees.

[https://pypi.python.org/pypi/codon\\_tools/0.2](https://pypi.python.org/pypi/codon_tools/0.2)

Toolkit to manipulate synonymous codon usage in various ways

<https://www.pytables.org/usersguide/tutorials.html>

To manipulate large objects in python hdf5

<https://www.h5py.org/>

Large file pythonic mode big data

<http://www.blopig.com/blog/2016/08/processing-large-files-using-python/>

Simple howto process large files

<https://lvdmaaten.github.io/tsne/>

Tsne better

<http://www.aklectures.com/lecture/calculating-isoelectric-point-of-proteins>

How to compute Pi

Usefull BOPython tools

<https://github.com/biopython/biopython/blob/master/Doc/doc.rst>

<http://biopython.org/DIST/docs/api/Bio.Graphics.GenomeDiagram-module.html>

[http://biopython.org/DIST/docs/api/Bio.Phylo.Applications.\\_Phyml.PhymlCommandline-class.html](http://biopython.org/DIST/docs/api/Bio.Phylo.Applications._Phyml.PhymlCommandline-class.html)

Parser/ ParserLike

<http://biopython.org/DIST/docs/api/Bio.SearchIO.FastaIO-module.html>

<http://biopython.org/DIST/docs/api/Bio.SearchIO.BlastIO-module.html>

<http://biopython.org/DIST/docs/api/Bio.SeqIO.InsdcIO-module.html>

<http://biopython.org/DIST/docs/api/Bio.SeqUtils.CodonUsageIndices-module.html>

<http://biopython.org/DIST/docs/api/Bio.SeqUtils.CodonUsage-module.html>

<http://biopython.org/DIST/docs/api/Bio.codonalign.codonalphabet-module.html>

<http://biopython.org/DIST/docs/api/Bio.codonalign.codonseq-module.html>

<http://biopython.org/DIST/docs/api/Bio.codonalign.codonalignment-module.html>

<http://biopython.org/DIST/docs/api/Bio.Data.CodonTable-module.html>

[http://biopython.org/DIST/docs/api/Bio.Align.Applications.\\_Muscle.MuscleCommandline-class.html](http://biopython.org/DIST/docs/api/Bio.Align.Applications._Muscle.MuscleCommandline-class.html)

<http://biopython.org/DIST/docs/api/Bio.pairwise2-module.html>



---

<http://biopython.org/DIST/docs/api/Bio.Alphabet.IUPAC.IUPACUnambiguousDNA-class.html>

<http://biopython.org/DIST/docs/api/Bio.Blast.NCBIWWW-module.html>

<http://www.stack.nl/~dimitri/doxygen/manual/starting.html>

doxygen

<https://github.com/Feneric/doxypypy>

Doxypypy

species, genes, entropyval+location+length

Unclassified ideas and things :

- you can present a 3D data and colorized similar genes in the same color and when hovering on a dot you display the species --kept
- 461 species \* 1000 genes -- changed
- are the entropy values most similar to the genes or the species ?
- you could do PCA for with each data points being species and values being the entropylocation\*codons
- But you only have 18 amino acids instead of 23 - 1. why ?  
Because you have 3 that are encoded by just one codon (no bias possible)
- Could we have an entropy location for each codons ( like a ratio inside a gene)

## Mars 2018

### Writting

---

MID MARCH

Writing a save and load function, writing a function for comparison and plotting of gene homology according to species.

Creating an explanation of my pipeline and goals for Tobias and Dominique to see it ( present in the readme ) \*requested by Dominique\*

Doing a refactoring of my code to fully match some logic.

Adding the ability to create group of homologies.

ENDMARCH

<http://rest.ensemblgenomes.org/> getting access to the API

We should try to plot many things right ?

- Distribution of homologies per species given phylo tree
- Entropy distribution of genes for each specie in the same order of genes and then fit a curve ( order according to distribution similarities ) look for KL divergence values.
- Entropy distribution total per species and compare to phylo tree

Really important , given the clusters and X species with the same set S of homologies, is there a relationship between the cluster

Else, not using the clusters, is there a relationship ( should be a similar transformation between their different Gene entropy vector )  $T(S(1,X(1))) \sim T(S(2,X(2)))$  etc..

## Avril 2018

MID APRIL

I am getting many feedbacks from tobias about my result and spend half a day discussing also with Yun and Dominique. I am now setting up my plan for the last months of development. The goal is work as much as I can on the next implementation of PYCUB and find the time to fully prepare myself for the exams.

-----

Goals

The goals of this project is to try to infer relations in the codon usage of different genes from different species using machine learning and to let researchers use it and improve it as they please.

---

A goal for me at this point is to have a the pipeline running.

In more long terms I wish to :

Pursue on the ideas defined here and on the TODO file.

I am also preparing a presentation available on slideshare soon.

And at the end of the summer, a Research Document.

#find more information in my INFO and TODO files.

## Pipeline

For now the Pipeline is able to :

- let me load and save the data in the right format,
- get the precomputed data from Kent University
- Visualize and cluster homologies
- Compute and visualize a k-mean clustering from a homology

Here is How the data-structure looks like :

\*to come\*

What happens exactly :

I am getting Yun's file directly from the dropbox server. Then I am mapping then to the datastructure and I am using currently the entropy-location values.

Each amino acid's entropy location values for one homology defines a dimension of a feature vector describing the codon usage statistics of a species' gene.

As it is required to do a good comparison, to have species that share a lot of homologies, I am clustering groups of species that the highest number of shared homologies. You can see this comparison as a 2D matrix showing binary values representing if a species has this particular homology. A similarity matrix can be displayed as well to see particular similarities.\*

Using each highly similar groups I am taking the homologies one by one and apply it another clustering algorithm on the entropy location vector to group species with similar entropy features. \*

---

Clusters for particular homologies can be visualized using T-sne Hinton.

<http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

To display them on a two dimensional surface (dimensionality reduction). This plot can be made interactive to see species corresponding to each dots

Using the statistical information on the process and data exploration on the outputted clusters, the next steps shall be defined.

Ideas revolve around :

- Trying to find explanatory statistics of the features that I have as an input using :
  - Autoencoders
  - Linear factor models
  - From the clusters using non linear classifiers that would explain some metadatas that Tobias is gathering.
- Using more data from other databases.
- Using other statistics on the codon usages.

\*repeated with different clustering algorithm and tested with statistical tests

## May 2018

BEGIN MAY

Working on the code mostly and continuing on investigating the many datasets I should need to use.

I have also started to work on the final exams.

### Things I will do for the last summer sprint of development:

- debugging of some parts of the code and adding more documentation
- Create the function to compute the phylogenetic distance matrix between genes from a pseudo half life of genes with the gamma function and their sequence similarity and see if there is a relationship with the CUB in any way.
- Plot as a long stacked barplot, for each species, how much its genes are outliers, how much are belonging to a secondary cluster and how much are belonging to the principal cluster.
- Compare the clusters together with a similarity matrix representing kind of a distance between clusters and do the same for the species

- 
- find if we are close to ancestry tree when doing our clustering
  - compute the F1 score (amount of species that should be in the cluster and are not, that should not be but are in the cluster etc...) regularized with respect to the size of each cluster, for each homologies.
  - set the parameters of the clustering algorithm to find clusters that best relate to ancestry distance
  - compute variance between clusters
  - look at the mean variance in CUB value and mean range for each homology
  - use a mean CUB values and compare it to all sequences values for one species and one homology
  - Use tRNA copy number to explain the variance in the data.(see how it explains the variance for each species. And which ones have their variances explained the most
  - Use entropy, entropy location, codon frequency, (normalized and unnormalized ones.), use random values as well (random values can be created from the same thing as the partition functions for entropy location.
  - Compute and compare averages over sequences/species/homologies
  - download or find more meta datasets and see what can be done with them  
[http://ensemblgenomes.org/info/data/protein\\_features](http://ensemblgenomes.org/info/data/protein_features) and  
[http://ensemblgenomes.org/info/data/repeat\\_features](http://ensemblgenomes.org/info/data/repeat_features) ( protein features and repeat features)
  - look at similarity distance in sequences and compare it to the CUB values I have whether similarity is entropy
  - Add ways to inspect the data (plottings, rankings...)

## **Juin 2018**

### **BEGIN JUNE**

Last face to face meeting with everyone. I have received advices and more direction of research and things to try. Overall, the idea was to continue on what I was trying to do. I had the confirmation I would receive the metadata. I received also some feedbacks from Yun which continued to explain me things I did not have fully understood about her work.

-----  
Most of the beginning month of June was also spent on the final exams.

### **END JUNE**

---

I have received the metadata from Tobias and integrating it to my pipeline. I had some issues with my computer which took me some time to resolve.

I also skipped Dominique about my research but did not have much to say as I was still developing the code and did not finish implementing everything.

-----  
I have spent 2 weeks on creating faster partition functions with relevant calculus on sheets of paper.

I am spending most of this month on the development of the code I am adding new functions and discovering more functions to add with time. I am spending a huge amount of time trying to better the measures that Yun has made and the partition that she is using.

I am also spending 20% of my day, reading research work on the codon usage bias, Machine learning and related topics.

## **July 2018**

MID JULY

I spend most of my time continuing coding PyCUB, I am in the phase of constant debugging and trying the function. I am not trying to get results yet. I will also try in the future to implement the work on 3D genomic distance if I have the time. I have however spent a week working on a different project. It slowed my progress but helped me think about something else and come back with more focus.

## **August 2018**

BEGIN AUGUST

Most of my time is spent debugging my software and adding some important functions that will be necessary to present more clearly my results. I am also gathering all the literature I can find and reading quickly related ones.

MID AUGUST

Skyping Dominique to have feedbacks about my results. Said I had made a lot of good work and that he will read and correct my work during the month. He also asked if I

---

would continue help them for a few more days. I would be tasked to retrieve some data and will be cited in the next paper to come.

We have also talked about how to write my thesis well and what should be the main information and level of description I should use in order to have a clear thesis.

-----

This entire month was about the writing of my thesis draft and debugging some final pieces of code (especially the one on Diament's Paper). (More about the process of writing my thesis can be found by looking at the timestamped log of my thesis draft here :

<https://docs.google.com/document/d/1FjMTWjNO9CuLI8HVdw1smfPYXc69yjsxggIIQv9o9c/edit?usp=sharing>

## September 2018

### BEGIN SEPTEMBER

I have skyped Tobias one last time to get feedbacks and ideas on how to interpret the results. However, with the little time left I was not able to integrate them into my PiPeline

----

Most of this week and a half was about finishing more of my thesis, getting some nice plots and gathering final results I could add in it. The work was very intense (approx. 12h per day) but a lot was done. I had my thesis read by different people and learned a lot about this type of work.

---

## More:

Unclassified ideas and things :

- you can present a 3D data and colorized similar genes in the same color and when hovering on a dot you display the species --kept
- 461 species \* 1000 genes -- changed
- are the entropy values most similar to the genes or the species ?
- you could do PCA for with each data points being species and values being the entropylocation\*codons
- But you only have 18 amino acids instead of 23 - 1. why ?  
Because you have 3 that are encoded by just one codon (no bias possible)
- Could we have an entropy location for each codons ( like a ratio inside a gene)

### *Important ideas*

" As with Euler's bridges, it's all about the connections. Social networks map out the relationships between people, with clusters of names (nodes) and connections (edges) illustrating how we're all connected. There will be clusters relating to family, college buddies, workplace acquaintances, and so forth. Carlsson thinks it is possible to extend this approach to other kinds of data sets as well, such as genomic sequences. "One can lay the sequences out next to each other and count the number of places where they differ," he explained. "That number becomes a measure of how similar or dissimilar they are, and you can encode that as a distance function." "

Quanta, december 2017

We could use graph theory to infer two things :

- The similarity manifolds of a group of gene/species nodes being species and links being homologous genes with distance being  $D = 1/\log(\#\text{homologues})$
- The group sharing the maximum number of homologous genes either by :
  - having a similar graph with notion of distances (weights), we would like to find the maximum clique (fully connected graph) with a \sum over their nodes' length that is small. Get only the closest nodes and don't go to far from the ref node.

<http://pub.ist.ac.at/~vnk/software.html>

**Blossom V: A new implementation of a minimum cost perfect matching algorithm.**



- 
- Get graphs for each homologies and compare them one by one, get the most similar ones and remove the differences.

### **Feature Correspondence via Graph Matching**

The GC content is influenced by many factors, for example the temperature as GC has more molecular strength than AT. (LINK2) As the codon bias could be explained by the GC content. However the GC bias receive the same problematic than Codon bias. It is the same problem on a different point of view.

### *Management*

I can have someone else working on the project but the problem is that they don't know more about biology or stats

He/She could be just really good dev and know data science, neural network, and more if there is a big requirement in the development side

Could help on scaling the code (for example doing the computation on a cluster or server)

I have send the project on the student forum.

## **Additional information**

More can be found looking at the github repository at <https://github.com/jkobject/pycub> . There, timestamped, readMe and TODO files are describing my thought process along the way.

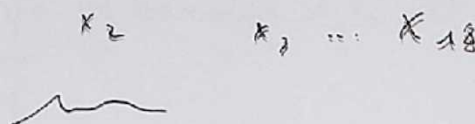
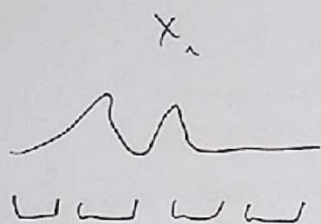
Moreover, many writings were done on paper. The most informative documents will be scanned and appended to this logbook

-----  
TOTAL NUMBER OF EMAIL EXCHANGED WITH RESEARCH GROUP: 302

NUMBER OF MEETINGS: ~13

NUMBER OF SKYPE: ~15  
-----

## **Additional Images/logs**



~~Value de x~~


~~Value~~ Value of constraint over the randomness of the distribution

if  $x$  means higher external constraint very likely

$x$  point in 18D space of the external constraint over each amino acids encodings for a particular gene

1. what can explain this constraint?

2. given a distribution of constraint, what factor can explain a variation of this constraint in any combination of the 18D

t-SNE gives you a manifold  represent

the variation of this constraint, given a set of species under  
for 1 homology,  
different external pressures.



Can it be explained by - size of the gene?

- similarity between sequences?
- evolution between species?



- are the doublets under same pressure?
- the length?
- the GC content?
- the 3D conformation distance of the DNA?
- a combination of these?

Do such computation make sense? (using random homology, do we find the same things?) (is it consistent amongst a group of homologues?)

now we can look at homologies. is there a relation between homologies in the average enthalpy forces on them? <sup>is it</sup> function, a ~~new~~ indicator of some sort of the entropy value? or is it <sup>length or GC content or ~~some~~ ~~they~~ ~~randomly~~</sup> ~~do they all cluster~~ ~~around the same low mean entropy cluster?~~

if

looking at each species is the constraint on their entire genome similar at the one on their CPS, or their homologous gene to cerevisiae?

do some of them have more global pressure than others?  
if so, what can explain it. tRNA pools?  
gene/genome sizes?  
activity?



and  
 $n = 3$

$$\Rightarrow \text{PMF}(\mathcal{W}(n_1, n_2, n_3, \text{len})) \frac{\text{len}!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3} = \frac{\text{len}! \times \left(\frac{1}{\text{len}}\right)^{n_1} \left(\frac{1}{\text{len}}\right)^{n_2} \left(\frac{1}{\text{len}}\right)^{n_3}}{n_1! n_2! (\text{len} - n_1 - n_2)!}$$

$n_1, n_2, n_3$

$$f(n_1++) \Rightarrow \frac{\text{PMF}(\mathcal{W}(n_1, n_2, \text{len})) \times \left(\frac{n_1+1}{\text{len}}\right)^{n_1+1} \left(\frac{\text{len} - n_1 - n_2}{\text{len}}\right)^{\text{len} - n_1 - n_2 - 1}}{(n_1+1) \times (\text{len} - n_1 - n_2 - 1) \left(\frac{n_1}{\text{len}}\right)^{n_1}}$$

0, 0, 0, 400

0, 0, 1, 399

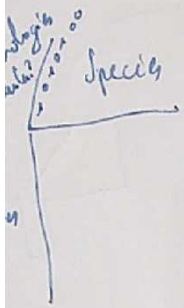
0, 0, 2, 398

$$P_{n_1+n_2+n_3} = p_1^{n_1} + p_2^{n_2} + p_3^{n_3}$$

$$\text{since } p_1 = p_2 = p_3 = \frac{1}{\#p}$$

$$\Rightarrow \text{PMF}(\mathcal{W}(n_1, n_2, \text{len})) = \frac{\text{PMF}(\mathcal{W}(n_1, n_2, \text{len})) (\text{len} - n_1 - n_2 - 1)}{n_1 + 1}$$

u B 28/66



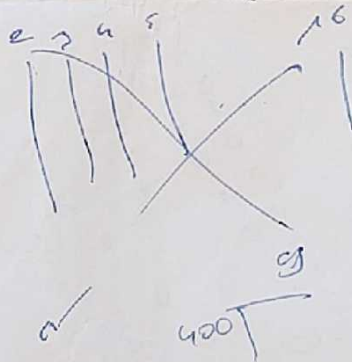
opens hooless array 30  
plot matrix per species

sum over ones / TOT  
similarity species  
species

28 matrix of ~~into~~ floaks

same with disks  
instead of clusters  
(in 180)

0  
2 2 6 14 18  
16 22 40 40



9 x 400  
2432  
3

18  
0 400

$$\frac{\log n \times \left( \frac{n_1}{\log n} \right)^{n_1} \left( \frac{n_2}{\log n} \right)^{n_2} \left( \frac{\log n - n_1 - n_2}{\log n} \right)^{\log n - n_1 - n_2}}{\log n! (\log n - n_1 - n_2)!}$$

for  $n_1, n_2, \log n$

for  $n_1, n_2, \log n$

valid =  $\log n! (\log n - n_1 - n_2)! \times \log n$

$$\frac{n_2 \times \log n - n_1 - n_2}{\log n}$$





$$\text{val } (i-1) \times \frac{n}{\log} \times (\text{len} - n - (n_2 + 1))$$

$$\frac{n \times (\text{len} - n - (n_2 + 1))}{\log}$$

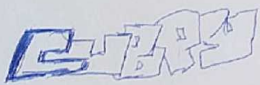
$$[x, y, z]$$

$$[1, 1, 0]$$

$$1 \ 1$$

$$0, 1, 1$$

$$0, 1$$



$$\frac{n!}{k! (n-k)!} \frac{3!}{2! (3-2)!}$$

$$\frac{\left(\frac{x+1}{a}\right)^x}{\left(\frac{x}{a}\right)^x}$$

$$\left(\frac{x+1}{x}\right)^x$$



$$\frac{x}{a}^x$$

$$\left(\frac{x+1}{a}\right)^{x+1}$$

$$\frac{x+1}{a} \left(\frac{x+1}{a}\right)^{x+1}$$

$$\left(\frac{x}{a} + \frac{1}{a}\right)^x$$

Can we neglect  $\frac{1}{a}$ ?  
if  $a \gg 0$

$$\frac{3!}{1! (2!)}$$

$$e^{x \ln(x)} (x+1) \ln(x+1)$$

$$e^{x \ln(x+1)} e^{\ln(x+1)}$$

