

# **Mutual Information Analyzer**

## **version 1.0.0**

**developed by: Flavio Lichtenstein**

**contact: flalix@gmail.com**

**Federal University of Sao Paulo**

**(UNIFESP - Universidade Federal de São Paulo)**

**DIS - Bioinformatics**

# Mutual Information Analyzer

## Summary

1.	Introduction .....	3
2.	Algorithms .....	3
3.	General visualization .....	5
4.	Species Listbox.....	5
5.	How to begin an analysis?.....	7
6.	Footer buttons.....	10
7.	Directories / Roots .....	11
8.	NCBI data acquisition .....	12
9.	Gbk to Fasta .....	14
10.	Alignment.....	16
10.1.	How the alignment algorithm works?.....	17
11.	Species List Box: operational .....	18
12.	Purging .....	19
13.	Consensus .....	21
14.	Vertical Mutual Information (VMI).....	22
14.1.	Vertical Shannon Entropy .....	26
14.2.	Vertical Mutual Information.....	27
14.3.	How to calculate Vertical Mutual Information.....	28
15.	Horizontal Mutual Information (HMI) .....	29
15.1.	Horizontal Mutual Information .....	32
15.2.	How to calculate Horizontal Mutual Information .....	32
16.	Jensen-Shannon Distance .....	33
16.1.	JSD from Vertical Entropy .....	36
16.2.	Jensen-Shannon Distance Definition.....	37
16.3.	JSD Standard Error - SE(JSD) .....	37
17.	Hierarchical Cluster .....	38
17.1.	Hierarchical Cluster Dendrogram .....	39
18.	Shannon Entropy (simulation) .....	40
18.1.	Shannon Entropy Simulation graphics .....	42
19.	Bibliography.....	43

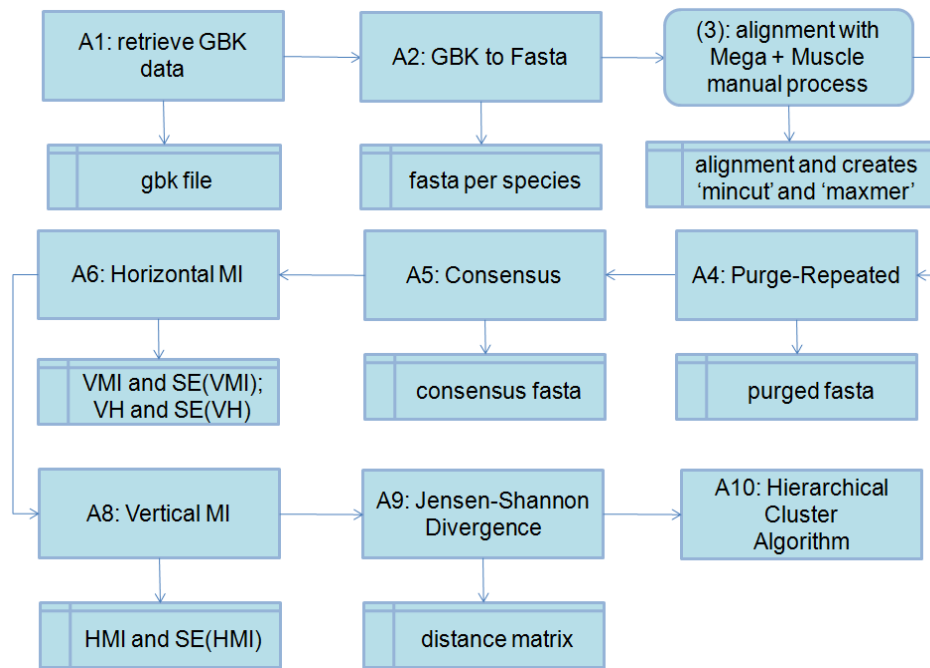
## 1. Introduction

Mutual Information Analyzer (MIA) is a pipeline written in Python (Biopython community, 2012) with the intent to retrieve, manipulate and store molecular sequences. Mia calculates Vertical Shannon Entropy, Vertical and Horizontal Mutual Information (VH, VMI and HMI, respectively). Once with VH, VMI and HMI distributions, Jensen-Shannon Divergence can be applied to calculate all distribution distances. Each pairwise species distribution distance and respective standard error are calculated and stored in Distance Matrices (ASCII files). These distances allow one to assess discrimination of different datasets (species distributions), and these data are displayed as histograms. They can also be clusterized, via a hierarchical cluster algorithm, and displayed as a dendrogram only with the intention of a better visualization.

## 2. Algorithms

Mutual Information Analyzer (MIA) is a pipeline with the following algorithms:

- A1) NCBI: gathers data in NCBI and stores them in GBK file format;
- A2) Gbk to Fasta: analyzes GBK file and organizes in fasta files per species;
- A3) Alignment: aligns sequences with Muscle (Edgar, 2004) and at the end creates two fasta files: "mincut" cutting out columns and sequences with large gaps, and "maxmer" maintaining the maximum possible gaps;
- A4) Purging: replaces ambiguous nucleotides via IUPAC nucleotide ambiguity table, and eliminates sequences with undesirable words in their names like "synthetic";
- A5) Consensus: replaces gaps by their vertical consensus nucleotide;
- A6) VMI: calculates and stores Vertical Entropy (VH) and Vertical Mutual Information (VMI) distributions, and plots the respective histograms and heat maps;
- A7) HMI: calculates and stores Horizontal Mutual Information (HMI) distributions, and plots the histograms;
- A8) JSD: calculates Jensen-Shannon Divergence, storing distances and their SEs in distance matrix files, and plots the histograms;
- A9) HC: calculates hierarchical cluster and present it as a dendrogram;
- A10) Entropy: simulates Shannon Entropy.



**Figure 1** – Mutual Information Analyzer pipeline

The first two algorithms are to get data (genbank files) from NCBI and split them into fasta files. More details are given below.

The third algorithm, Alignment, cut out "columns" and "sequences" with too many gaps. These actions results in two distinct fasta files: "mincut" where minimum length sequence means that only few gaps were allowed and "maxmer" where maximum length sequence means that a little more tolerance were given to gaps that were found.

VH (Vertical Shannon Entropy), VMI (Vertical Mutual Information) and HMI (Horizontal Mutual Information) are the next algorithms, and calculate these distributions for "mincut" and "maxmer". Mia also apply a bias corrections (Roulston, 1999) if wanted, a correction necessary for limited length sequences. Once with "mincut" and "maxmer" Mia allow assess of the gain or loss of information, with or without bias correction.

Distances between distributions were calculated via the square root of JSD. Since JSD is not linear function of the data their standard errors are calculated by empirical propagation.

### 3. General visualization

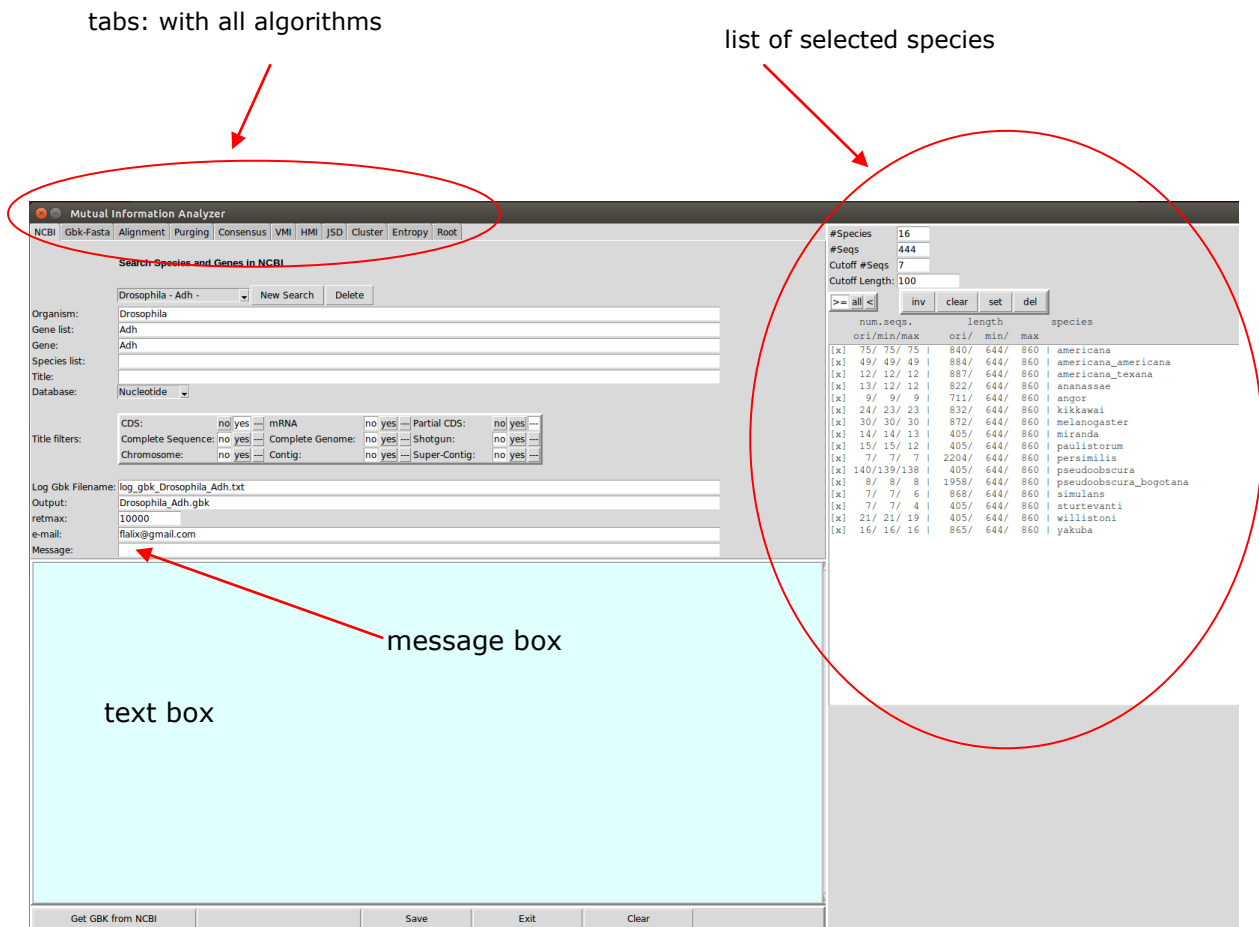


Figure 2 – General visualization of the first tab of MIA and Species List box at right

### 4. Species Listbox

At the end of the second algorithm Species Listbox will be filled with all species that you search could found. But, pay attention, many Species in NCBI have 1 to 5 sequences available. That is too few, since MIA needs at least 7 to 10 sequences to calculate VMI and VH with low SE. Therefore we should define a limit, e.g. at least 7 sequences. If 7 is chosen then a few sequences should appear, otherwise if you choose 15 more sequences should appear.

Another filter is Cutoff Length: that is the minimum length expected for the selected sequences. Some studies in NCBI present short sequences, really very short sequences, e.g. studying promoters or active site regions. These sequences must be ignored. If you are more exigent you can define the Cutoff Length equal to 400 or more, depending the gene length that you are to studying.

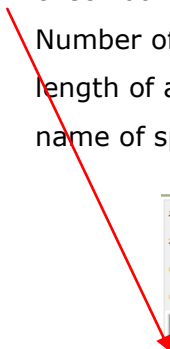
Once defined the Cutoff Number (#) of Sequences and Cutoff Length, you may continue your analysis and processed files will have these numbers in their filenames (see input/output filenames below).



Thus, Cutoff Number (#) and Cutoff Length define an unique experiment. If you change one of these values you must restart your whole study from the Alignment Algorithm.

The Species Listbox header:

- check box (double click to set or reset)
- Number of found sequences: original, mincut, maxmer
- length of aligned sequences (L): original, mincut, maxmer
- name of species from a given Organism (in this case: Drosophila)



		num.seqs.			length			species
		ori/	min/	max	ori/	min/	max	
#Species	16							
#Seqs	444							
Cutoff #Seqs	7							
Cutoff Length:	100							
		<input type="checkbox"/> >= <input type="checkbox"/> all <input type="checkbox"/> < <input type="button" value="inv"/> <input type="button" value="clear"/> <input type="button" value="set"/> <input type="button" value="del"/>						
[x]	75/ 75/ 75	840/	644/	860				americana
[x]	49/ 49/ 49	884/	644/	860				americana_americana
[x]	12/ 12/ 12	887/	644/	860				americana_texana
[x]	13/ 12/ 12	822/	644/	860				ananassae
[x]	9/ 9/ 9	711/	644/	860				angor
[x]	24/ 23/ 23	832/	644/	860				kikkawai
[x]	30/ 30/ 30	872/	644/	860				melanogaster
[x]	14/ 14/ 13	405/	644/	860				miranda
[x]	15/ 15/ 12	405/	644/	860				paulistorum
[x]	7/ 7/ 7	2204/	644/	860				persimilis
[x]	140/139/138	405/	644/	860				pseudoobscura
[x]	8/ 8/ 8	1958/	644/	860				pseudoobscura_bogotana
[x]	7/ 7/ 6	868/	644/	860				simulans
[x]	7/ 7/ 4	405/	644/	860				sturtevantii
[x]	21/ 21/ 19	405/	644/	860				willistoni
[x]	16/ 16/ 16	865/	644/	860				yakuba

**Figure 3** – Species List box and their command buttons

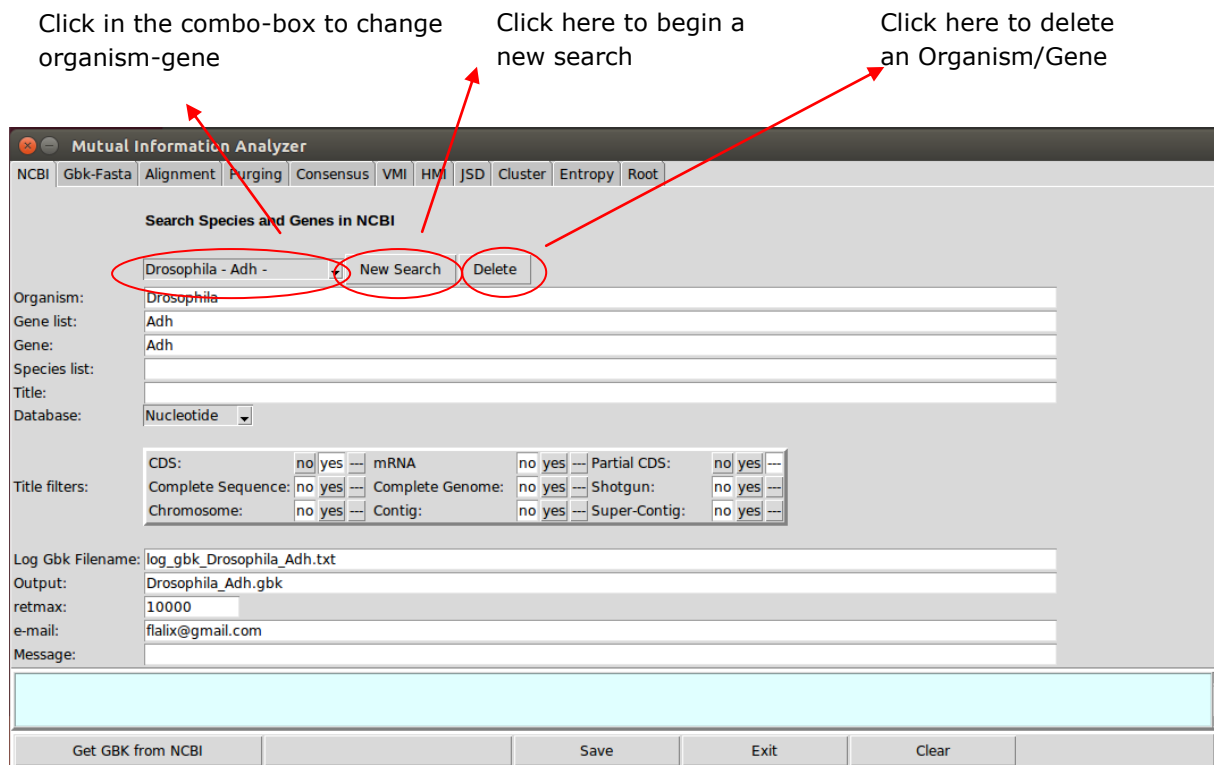
In the alignment tab (3rd) you may click in one of the lines of the Species Listbox and choose a determined species. After this you may see the data at Seaview (Gouy et al., 2010) clicking in <Seaview: species>, if it is installed in your computer. You also see at left two buttons, one to see mincut aligned sequences and other to maxmer aligned sequences.

## **5. How to begin an analysis?**

To initiate a study you must be in the first tab called NCBI. MIA in its first version only get data from NCBI (<http://www.ncbi.nlm.nih.gov/>). All analyses explained bellow can be reproduced at NCBI site prompt.

The basic steps are:

- click in "New Search", then the screen is cleaned shortly thereafter
- define one Organism and one Gene and/or a Title (better one Gene only, first)
- some genes have many isoforms, therefore, in these cases, you may define a list of genes (a string, comma separated) including them or excluding. Thus, you may define a Species list (a string, comma separated). This list is optative.
- Title: may be a useful filter to find one or many words via Genebank Title registers. As seen bellow, all filters (CDS, mRNA ... Super-Contig) also use Title to filter some conditions.
- Click on <Get GBK from NCBI> button and MIA will find all species for a determined Organism/Gene via NCBI search engine. Or you may choose more filters, see below.



**Figure 4** – MIA first tab, interfaces NCBI via a BioPython API.

Once defined an Organism and its Gene, you should define if you want to study a DNA Sequences (Nucleotides) or Amino Acids Sequences (Proteins).



In this version you may acquire both, but MIA will only analyze Nucleotides (DNA) at version 1.0.0.

It is obvious that there are many different analyses for:

- a) only one gene
- b) many genes
- c) 2 kpb<sup>1</sup> before and 2 kpb after the gene
- d) mRNA (one or many)
- e) compare information between whole DNA segment or cDNA
- f) proteins (one or many)

---

<sup>1</sup> kbp - kilo base pair



In this version MIA is prepared to run, automatically, option a, "only one gene". But you may acquire a gene sequences with many pair bases (pb) before the gene and many pb after it (option c). You will be looking for interaction sites and conservative positions in the intergenic regions. This must be a manual operation, in this version, and when saving files the names must be MIA compliance (see input/output filenames).

The next step is additional filters. The options are: CDS, mRNA, Partial CDS, Complete Sequence, Complete Genome, Shotgun, Chromosome, Contig, Super-Contig. For each filter option you may choose "No" (don't want), "Yes" (wanted, must exist in the title), "--" (indifferent, don't care). These options are useful to study CDS and partial CDS without dealing with a whole genome, because this may have millions to billions of base pairs.



"mRNA" should be avoided because these sequences may come only with the merged exonic regions (cDNA) and in your analysis you are searching for Mutual Information and don't want to lose a lot interactions.

Other fields:

- Log filename is the log resulting from your defined search. The search is "fault tolerant", that means, each data retrieved from NCBI (gbk) is stored in your own computer and registered in the log. If the connection is broken, in the next run MIA will skip all data acquired and ask NCBI only for the new ID's.
  - Output file: is the Genbank file resulting from your defined search.
  - retmax: is the maximum sequences that may be retrieved from NCBI. Let this number very high, e.g. 10,000.
  - e-mail: is your email. To ask for some service at NCBI you must identify yourself.
  - Message: running any one of the algorithms they will echo messages in the Textbox (see Figure 2) and each message will quickly appear in Message Box. If anything goes wrong look at these boxes.
- 
- **Delete:** this option (button adjacent to <New Search>) deletes the study from the combo-box. But all data will be preserved in your hard drive.



When you delete a study, MIA does not delete the directories where the data are stored. You must go there and delete yourself. See "roots" to find the path.

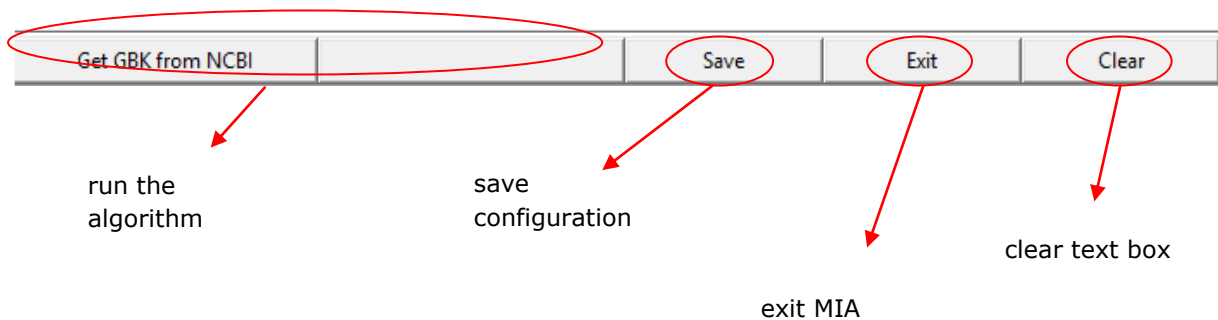
## 6. Footer buttons

On the footer we can see five buttons (Figure 5). The first two are to run algorithms. In some cases like in VMI-tab we will see two options (e.g. run Vertical Entropy and the other to run Vertical Mutual Information ).

The next button is Save. It saves all configurations, all defined options.



You should save the configurations after defined a new search. Or in any tab after changing the defaults, if wanted.



**Figure 5** – At the bottom we find run, save, exit and clear command buttons.

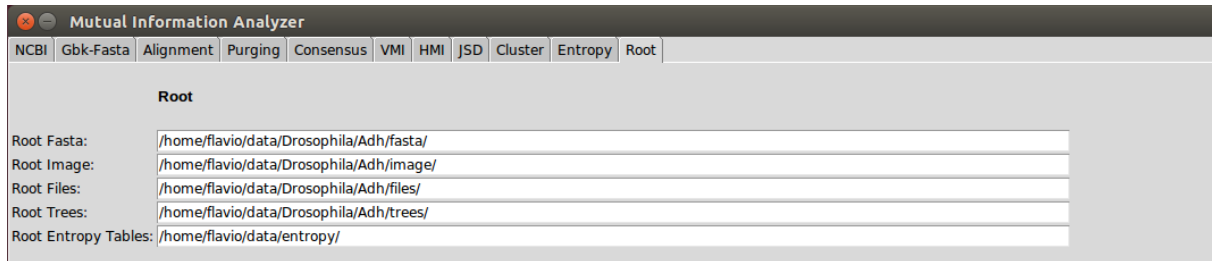
Clear: it cleans the Text Box.



All data echoed in the Text Box can be copied and pasted to a text editor.

## 7. Directories / Roots

The Root tab shows the directories where the data is stored.



**Figure 6** – The last tab shows the roots where fasta files, data tables, configurations, logs and images can be found

- In Windows it looks like: "c:\Users\Your\_Name\data\ ...."
- In Linux it looks like: "~/Your\_Name/data/ ...."

For each study (MIA analysis) you must define an Organism and a Gene. Confirming your parameters, you must save them clicking in <Save>. Then MIA will create many subdirectories like:

- home/Your\_Name/data/Your\_Organism/Your\_Gene/fasto → for fasta files
- home/Your\_Name/data/Your\_Organism/Your\_Gene/image → for images (histograms, heat maps, dendrograms, etc).
- home/Your\_Name/data/Your\_Organism/Your\_Gene/files → for calculated data files (calculated parameters, vmi data, hmi data, etc).
- home/Your\_Name/data/Your\_Organism/Your\_Gene/trees → where you may store phylogenetic trees obtained from softwares like Mega, Mr.Bayes .... (optional)
- home/Your\_Name/data/entropy/ → where Shannon Entropy tables are stored if you do this simulations.

## 8. NCBI data acquisition

Algorithm 01 (A1), tab NCBI, is the data acquisition algorithm. It is a module that interacts with NCBI Nucleotide/Protein database, via an API, retrieving Genbank data after the definition of at least one Organism and one Gene as parameters. Gene List and Species List are optional. Title may be defined to complement the filters or to substitute Gene parameter (you may define only a Gene, only a Title or both).

MIA is prepared to work only with Nucleotides in its first version, because it is much more informationally efficient.



you may acquire Protein data, but you will not continue the analysis because it is not implemented in version 1.0.0.

**Figure 7** – How to get data from NCBI - in genbank format

Title Filters are pre-prepared filters. The options are: CDS (coding sequences), mRNA, Partial CDS, Complete Sequence, Complete Genome, Shotgun, Chromosome, Contig and Super-Contig, which are strings to be searched (or not) at the register titles.

For each filter you may choose "No" (don't want), "Yes" (wanted, must exist in the title), "---" (indifferent, don't care). This option is useful to study CDS and partial CDS without dealing with the whole genome, because this may have millions to billions of base pairs.

As you define your parameter the Log Gbk filename and Output Gbk filename are defined. You cannot change them. Once defined, confirm "retmax" (maximum return of registers) and you "e-mail". At this point you may <Save> your configuration.

Right after click in <Get GBK from NCBI> and, via BioPython (Biopython community, 2012) API, data will begin to appear in the Textbox. If something wrong happens, don't worry (e.g. communication break, energy break, etc.), restart MIA and click again in <Get GBK from NCBI>. It is fault tolerant, and will continue at the point that communication broke.

Output file:

- Output: Gbk file, like "Drosophila\_Adh.gbk" (organism + gene)

## 9. Gbk to Fasta

Algorithm 02 (A2), tab Gbk-Fasta parses the Gbk file and saves on a fasta files. Before starting this step, be careful, define "Cutoff Length", "Maximum Length Sequence" and "Minimum Length Sequence". The first of these three parameters is used to define the output filename (see below) and must be maintained until the end of the analysis.

NCBI Gbk-Fasta Alignment Purging Consensus VMI HMI JSD Cluster Entropy Root

**Transform GBK to Fasta**

Organism:

Gene:

Title:

Cutoff Length:

Maxi length seq:

Mini length seq:

☐ Stop ☐ Show Message

Input:

Output:

Message:

Split Gbk into Fasta Save Exit Clear

**Figure 8** – Second Tab, Gbk-to-Fasta, splits gbk file in fasta files, one for each species before alignment

At the end of processing, the gbk-to-fasta file slices the gbk file in little fasta files, one for each species. These are the files that will be aligned in the next algorithm.

Input / Output files:

- Input: gbk file, like "Drosophila\_Adh.gbk" (organism + gene)
- Output: fasta file, like "(organism)\_(species)\_(type)\_Adh.fasta"; one file per found species.
  - (organism) is the organism name
  - (species) is the species name
  - (type) is:
    - Gene, CDS, Exon, Intron or Protein
    - This algorithm creates five files for each species. Here, as observed, the most complete dataset is Gene and we suggest that the option "Gene" should be maintained in whole analysis.

This algorithm also fills the Species Listbox with all sequences found.



Define a "Cutoff Number of Sequences" and spend a time observing the filters ">=", "all", "<". With ">=" only species that have number of sequences greater equal than "Cutoff # of Sequences" will appear, and "all" and "<" shows all species and less than, respectively.

## 10. Alignment

Algorithm 03 (A3), tab Alignment, is the alignment step. Here MIA aligns the sequences automatically via Muscle (Edgar, 2004), or you may align manually observing the output filenames compliance. In this first version we only provided Muscle algorithm. If you don't like it for any reason, do it manually or write to us.

NCBI Gbk-Fasta **Alignment** Purging Consensus VMI HMI JSD Cluster Entropy Root

**Align Muscle or Clustalw**

SeaView: mincut SeaView: maxmer SeaView: species

Organism: Drosophila

Gene: Adh

Title:

Species:

Cutoff #Seqs: 10

Cutoff Length: 100

Method: Muscle

Max Horizontal gaps: 30 %

Max Vertical gaps1: 10 % low value creates mincut

Max Vertical gaps2: 20 % high value creates maxmer

☐ Realign each species ☐ Realign all together

Input: Drosophila\_(species)\_(type)\_Adh\_100L.fasta

Output: Drosophila\_(max/min)\_Gene\_Adh\_100L\_cutoff10\_aligned.fasta

Message:

Align Save Exit Clear

**Figure 9** – Third Tab, Alignment, aligns each fasta file and at the end merges all them and realigned all sequences together. All this alignments can be seen in Seaview. At the end of this algorithm "mincut" and "maxmer" fasta files are created.

Parameters must be carefully defined. "Cutoff Length" should no more change and "Cutoff Number of Sequences" must be also well established.



Percentage filters, must defined:

- wished minimum percentage of gaps for a horizontal cutoff.
- wished minimum percentage of gaps defines "mincut" sequences,
- wished maximum percentage of gaps vertically defines "maxmer" sequences.



Observing maxmer, many gaps may appear in each sequence. This allows "minimum percentage of horizontal gaps" to delete many sequences, resulting in less sequences in each species. Thus this parameter must be high or at least "flexible". Study its behavior for each study case.

Input / Output files:

- Input: fasta file name like "Drosophila \_(species)\_(type)\_Adh.fasta"
- Output: 2 fasta files per species (for particular 100 of length and 7 for minimum number of sequences per species):
  - "Drosophila\_maxmer\_Gene\_Adh\_100L\_cutoff7\_**aligned**.fasta"
  - "Drosophila\_mincut\_Gene\_Adh\_100L\_cutoff7\_**aligned**.fasta"
  - where:
    - Drosophila is the organism
    - maxmer or mincut
    - Gene (for Gene, CDS or Exon)
    - Adh for the gene
    - 100L - minimum Cutoff Length = 100
    - cutoff7 - minimum number of sequences = 7

### 10.1. How the alignment algorithm works?

First MIA aligns each species sequences, thus you can observe them. Afterwards it merges all aligned sequences in a unique fasta file and aligned again. This is an important fasta file, with all sequences aligned, otherwise we could not compute and compare mutual information.

In this moment what do we have? A merged fasta file with a lot of gaps.

And how to find a consensus to analyze Entropy and Mutual Information? - gaps must be removed, cutting vertical positions full of gaps and next cutting out a bunch of sequences

full of gaps. At the end, in another algorithm vertical positions are replaced by their consensus nucleotide.

Therefore the next step is to maintain all vertical positions that have at least a "maximum desired percentage of gaps". If these columns exceed this quantity they will be deleted. You must define two limits: Maximum Vertical Gaps 1 and 2. The first is a low value and results in "mincut" fasta file, and the second must have a higher value, allowing more gaps, resulting in the "maxmer" fasta file.

Once in possession of "mincut" and "maxmer" fasta files MIA search for sequences that exceed the "maximum percentage of horizontal gaps". In other words, if one sequence has more gaps in than allowed MIA deletes it. Thus, each line is analyzed to assess if it will be retained or deleted. Maxmer has a greater probability to have sequences deleted.

At the end, all sequences have the same length  $L_{\min}$  for mincut files and the  $L_{\max}$  for maxmer files.  $L_{\min}$  and  $L_{\max}$  may differ from the original  $L$ , depending on the cutoff parameters. The number of sequences may be altered too, because some sequences presenting too many gaps were deleted from the dataset.

It must be emphasized that MIA cuts vertically positions with many gaps. Once we look for discrimination of closely related species, is expected that deletion of many gaps columns should not create a large informational difference to conserved genes from eukaryotes. Another concept that must be highlighted is that after cutting sequences the frame could be incorrect in terms of the reading frame. But, to Entropy and Mutual Information this important concept is not necessary in our analysis. Of course, a correct reading frame alignment and the conservation of all correct residues like in nature are much better choices. But little changes will not result in drastic differences over Vertical Entropy, Vertical Mutual Information and Horizontal Mutual Information.

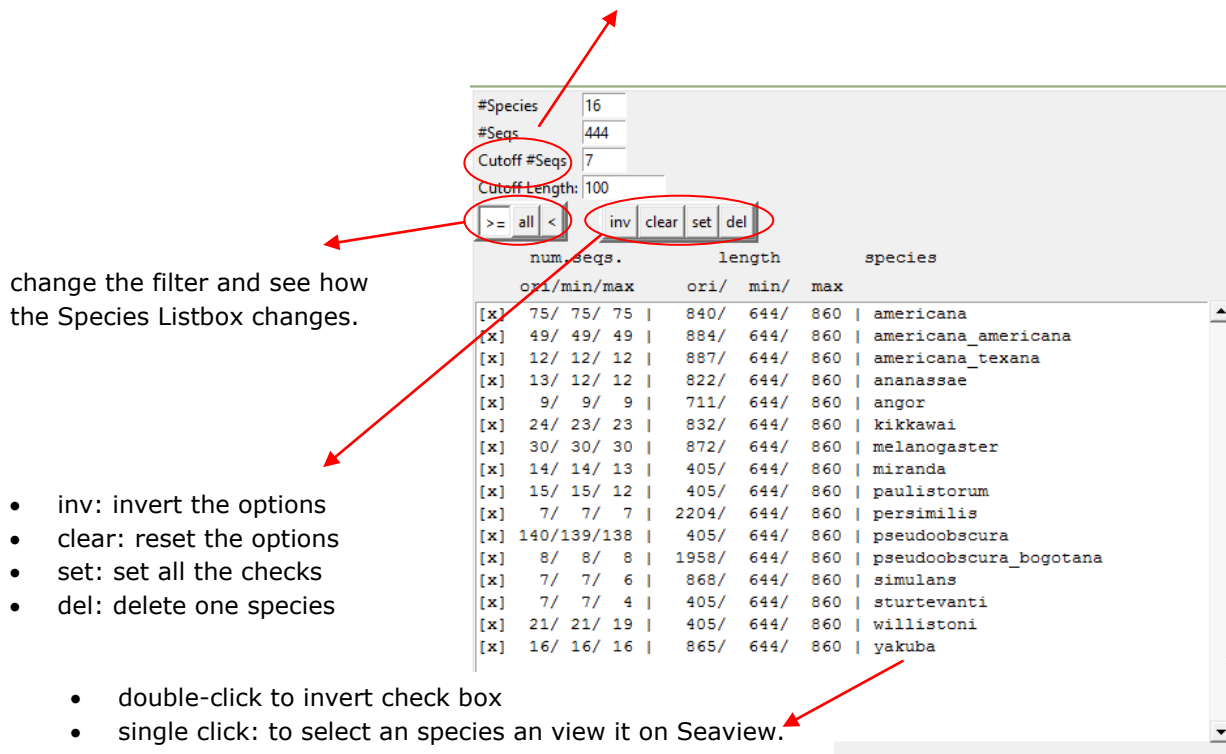
## **11. Species List Box: operational**

For the purpose to view, analyze and select species we built the Species Listbox. Thus try:

- change Cutoff #( Number) of Sequences
- click on ">=", "all" or "<"

and see how many species you have selected,

define Cutoff number(#) of Sequences



**Figure 10 – Species List Box**

## 12. Purging

Algorithm 04 (A4), tab Purging, is an internal sequence filter step. It looks for all positions in both set of sequences, mincut and maxmer, searching for nucleotides different than {A,G,T,C,-}. It replaces ambiguous nucleotides via IUPAC nucleotide ambiguity table (Johnson, 2010) with the most consensus nucleotide.

It also looks for "strange" words in the description (Title) of the sequence (fasta). For instance, we don't want to analyze synthetic data, therefore "synthetic" is one of the filter keys. At the end all synthetic sequences are deleted. Other terms may be include, just write them separated by comma.

Finally, the check box "Confirm Gene" tells MIA to only accept sequences with gene equal the Gene Parameter. This is a very rigorous restriction. Better should leave this check box off.

"Show message" check box will turn on "much more echo warning messages", leave it off.

The alignment algorithm creates Gene, CDS, Exon and Protein fasta files. Choose "Gene" because it is the most completed fasta file comparing to CDS and Exon. In the next version we want to change these parameters to Gene, periferic+Gene and Protein.

**Figure 11** – Fourth Tab, Purging, search for ambiguous nucleotides and bad sequence description

- Input: 2 fasta files per species
  - Drosophila \_maxmer\_Gene\_Adh\_100L\_cutoff7\_aligned.fasta
  - Drosophila \_mincut\_Gene\_Adh\_100L\_cutoff7\_aligned.fasta
- Output: 2 fasta files per species
  - Drosophila \_maxmer\_Gene\_Adh\_100L\_cutoff7\_**purged**.fasta
  - Drosophila \_mincut\_Gene\_Adh\_100L\_cutoff7\_**purged**.fasta

## 13. Consensus

Algorithm 05 (A5), tab Consensus, is a gap replacer. It looks for positions in both set of sequences, mincut and maxmer. For each vertical position (residue) it looks for gaps. Finding one or more, MIA calculates the consensus and replaces all gaps with it. As we may notice, this action introduces information.

In this version we did this simplified replacer. But consensus is a discrete distribution and we should replace proportionally at random based on found column polymorphism.

The "start and stop codon" check-box tell MIA to look for Start Codon ('ATG') and Stop Codons ('TAA', 'TAG','TGA')<sup>2</sup>. If you are not sure on your alignment and how MIA cut your sequence better leave this check box off.

The screenshot shows the 'Consensus' tab in the MIA software interface. The top menu bar includes 'NCBI', 'Gbk-Fasta', 'Alignment', 'Purging', 'Consensus' (selected), 'VMI', 'HMI', 'JSD', 'Cluster', 'Entropy', and 'Root'. The main panel is titled 'Substitute Ambiguous Nucleotides'. It contains a dropdown menu for 'Organism' set to 'Drosophila - Adh -', a text field for 'Gene' set to 'Adh', a 'Cutoff #Seqs' field set to '7', and a 'Cutoff Length' field set to '100'. Below these are three tabs: 'Gene' (selected), 'CDS', and 'Exon'. There are two checkboxes: 'Start and Stop Codon' (unchecked) and 'Show Message' (unchecked). The 'Input' field shows 'Drosophila\_(max/min)\_Gene\_Adh\_100L\_cutoff7\_purged.fasta' and the 'Output' field shows 'Drosophila\_(max/min)\_Gene\_Adh\_100L\_cutoff7\_consensus.fasta'. A 'Message' field is empty. At the bottom, there is a large light blue area for output and a footer bar with buttons: 'Consensus / Split species', 'Save', 'Exit', and 'Clear'.

**Figure 12** – Fifth Tab, Consensus, replace gaps by vertical consensus nucleotide

Input / Output files:

<sup>2</sup> For nuclear eukaryotes. For mitochondria and not prokaryotes these Stop Codons are not correct.

- Input: 2 fasta files per species:
  - "Drosophila \_maxmer\_Gene\_Adh\_100L\_cutoff7\_purged.fasta"
  - "Drosophila \_mincut\_Gene\_Adh\_100L\_cutoff7\_purged.fasta"
- Output: 2 fasta files per species:
  - "Drosophila \_maxmer\_Gene\_Adh\_100L\_cutoff7\_**consensus**.fasta"
  - "Drosophila \_mincut\_Gene\_Adh\_100L\_cutoff7\_**consensus**.fasta"



If you aligned your sequences manually the final results must have the name like the input filenames seen above, and then you run the consensus algorithm to fill the gaps, our they must have the output compliance filename and you can start VH, VMI and HMI.

## 14. Vertical Mutual Information (VMI)

Algorithm 06 (A6), tab VMI, is the VMI algorithm that calculates Vertical Entropy (VH) and Vertical Mutual Information (VMI). Once there are four possibilities:

- Mincut:
  - without bias correction
  - with bias correction
- Maxmer:
  - without bias correction
  - with bias correction

we decide to help the user with the button <calc all> (see below). That means, if you click <calc each> MIA will obey your choice of Mincut or Maxmer and observe if "bias correction" check box is on or off. Once each calculation may spent a lot of time (from a few minutes to many minutes depending on how large your sequence is, how many sequences you presented and the Number of Letters (NOL) chosen) we decided to create <calc all> and let MIA do all this job.

**Figure 13** – Sixth Tab, VMI, calculates VH and VMI and display their graphics

#### Options:

- "organism": the desired organism (read-only)
- "gene": the desired gene (read-only)
- "min/max": choose between mincut or maxmer sequences
- "calc each"/"calc all": calculates each chosen screen parameters, or calculates all four options ("min and max" versus "without/with bias correction").
- "number of letter": you can vary NOL from 1 to any other integer (recommended < 10).
- "Gene/CDS/Exon": recommended Gene (assess your data).

#### Other options:

- "Save Data": saves "summary" file at the end of all calculations. Summary file is a summary in a table format plus the ANOVA result.



Once you end the whole job we recommended turn this flag off, and if possible do a backup.

- "See image": displays the resulting image (a distribution, a histogram, etc.). Be careful that there are so many images equal to the combinatory number of selected options (species versus min/max versus bias correction, this can be a number like  $15 * 2 * 2 = 60$  images!). That can be dozen of images that you must spend a time analyzing and closing. If you close MIA, all images will be closed too. Better save all without seeing and then choose one or two to see or open them in the operational system.
- "Save image": if checked saves image in its directory (e.g., `~/data/Drosophila/image`) and you can choose which type (png, tif and jpg) and the resolution (in dpi, we recommended 120 for draft, 300 for a good resolution and 600 for a very good resolution).
- "Recalculate": once calculated, MIA will not do the job again, unless you turn this check box on.
- "Normalized": if checked divides all Entropy or Mutual Information values by the number of letters, normalizing them. Let this button initially off to observe how Entropy and MI vary. If you turn off and after turn on, MIA will not recalculate all values again, is a fast operation.
- "bias correction": apply Roulston bias correction
- "mnat": once most of the calculation results in values less than 1 nat, this option multiplies each value by 1000 given in mnat (millinat) unit.



In Widows when choose see image and save image, for a large number of images sometimes occur "out of memory" in an internal function of Matplotlib. In Linux we didn't see this problem.



Vertical Entropy is presented as a large histogram (Figure 14) that can be normalized giving rise to a distribution. But Vertical Mutual Information is a heat map (Figure 15). Here we present the Heat map in two dimensions (2D) or three dimensions (3D) (Figure 16). The options are:

- "2D heat map apply ceil": apply a ceil to all heat maps. This is good to compare all heat maps with the same maximum. But, is very bad if some of them are flat with low values and you are not able to distinguish the colors (everything is very blue).
- "ceil": the maximum you want in nat or mnat
- "3D": Rather than 2D you want 3D heat map. There is an interface in Python that enables the user to rotate images (we provided this function only in Windows, but we utilized a fixed frame in Linux). The 3D graphic is divided in 5 color grades to a better see the picture.
- "color scheme": many different color schemes that the user may choose.

Run VMI:

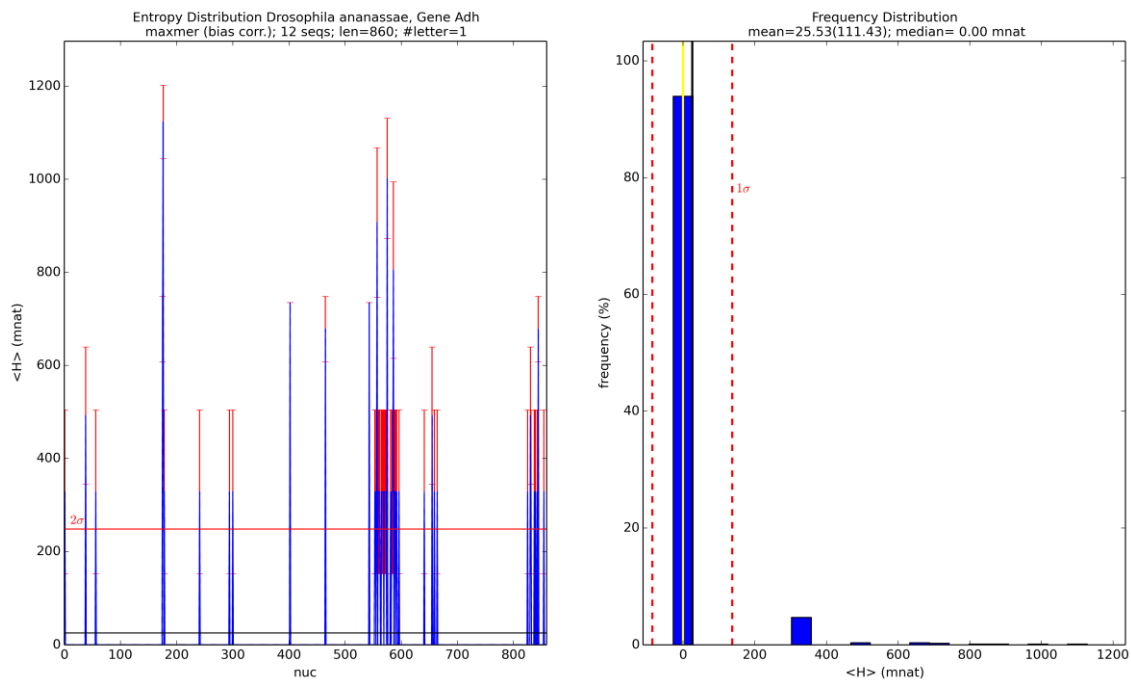
- click on <Vertical Entropy> to calculate and display Vertical Shannon Entropy.
- click on <Vertical Mutual Information> to calculate and display VMI
- click in anyone after choosing <calc all>

Input / Output files:

- Input: 2 fasta files per species:
  - "Drosophila \_maxmer\_Gene\_Adh\_100L\_cutoff7\_consensus.fasta"
  - "Drosophila \_mincut\_Gene\_Adh\_100L\_cutoff7\_consensus.fasta"
- Output: many files in dictionary python format:
  - All four possibilities are calculated: (mincut or maxmer) x (without correction or with correction)
  - VMI\_Drosophila\_min/max\_(species)\_Gene\_(gene)\_NOL(x)\_100L\_cutoff7\_(type)(bias).txt
  - where:
    - gene: like ADH, AMY, etc.
    - NOL(x) = number of letters: word size (e.g. from 1 to 7)
    - type:
      - \_hShannon: for Shannon entropy

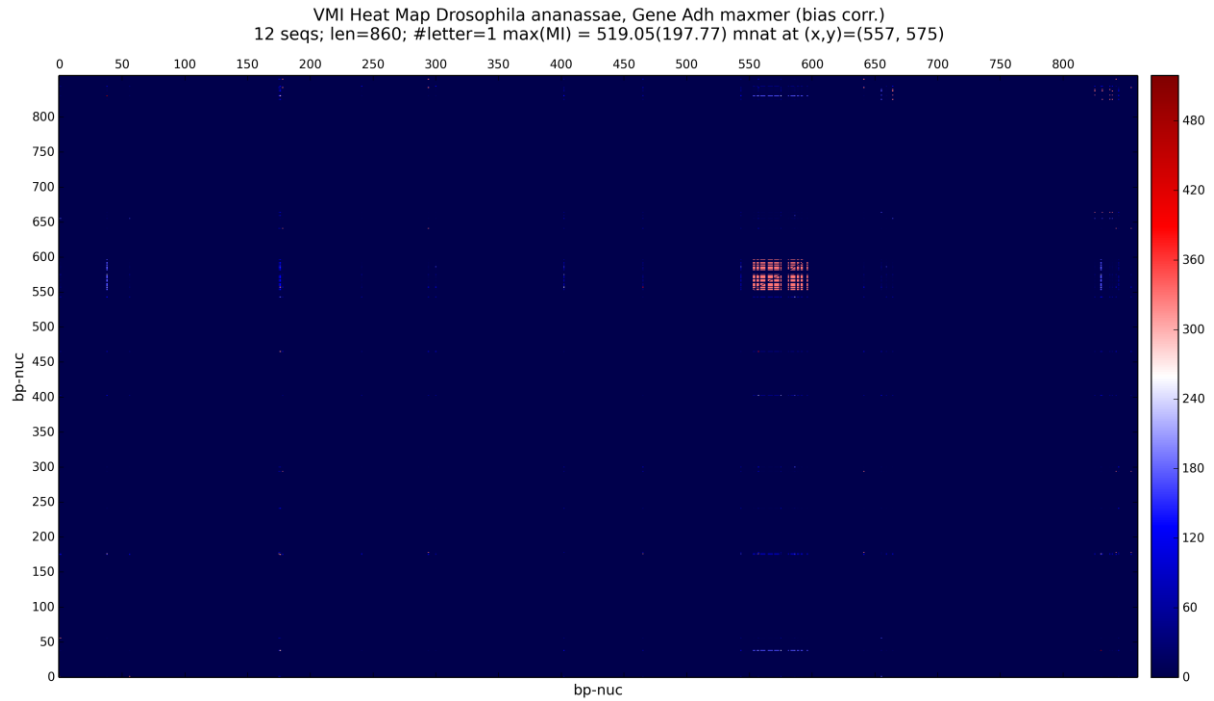
- `_seh`: Standard Error from HShannon
- `_pij`: nucleotide probabilities
- `_mi`: mutual information (MI)
- `_se`: Standard Error from MI
- `bias`: "`_bias_corr`" for bias correction or nothing without bias correction.
- `VMI_params_Drosophila_min/max_Gene_(gene)_NOL(x)_100L_cutoff7_(bias).txt`
  - "`params`" is a summary table with all information.

## 14.1. Vertical Shannon Entropy

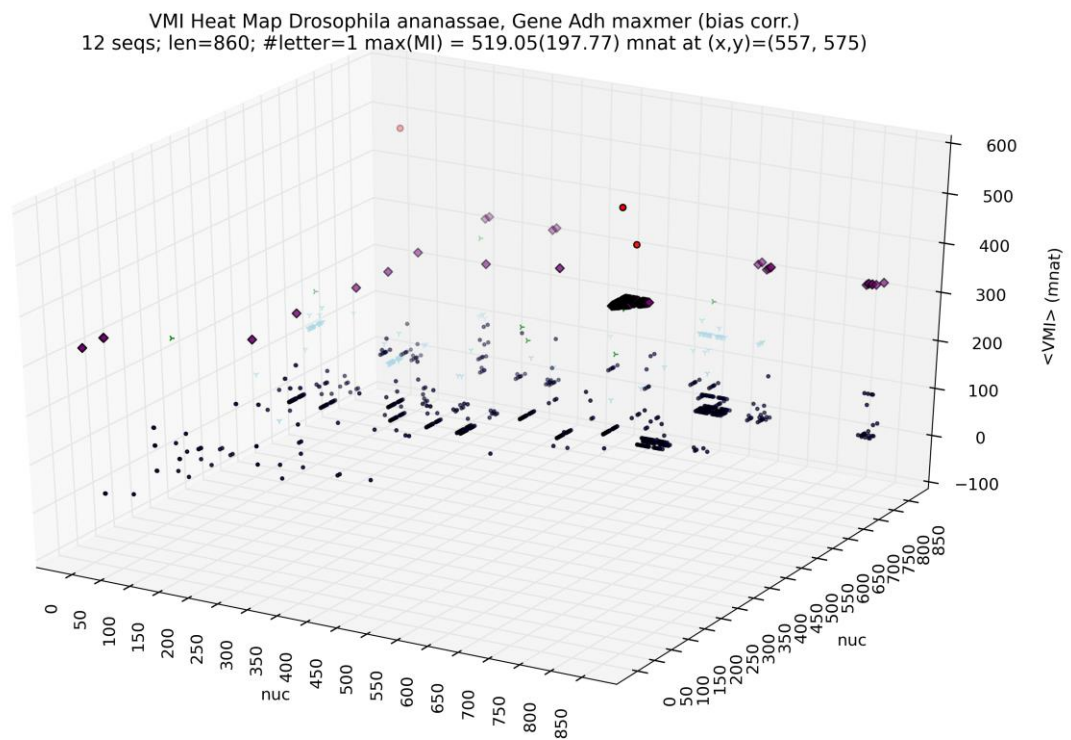


**Figure 14** - At the left the VH graphic – *Drosophila ananassae* 12 sequences, L=860 nucleotides, NOL=1, Gene Adh, maxmer with bias correction - horizontal red line stands for 2 standard deviation of the distribution the black line is the mean. At the right we see the frequency distribution graphic with 4 vertical lines: the two red standard deviation, in yellow the median and in black the mean of the VH dataset.

## 14.2. Vertical Mutual Information



**Figure 15** - VMI 2D heat map – *Drosophila ananassae* 12 sequences, L=860 nucleotides, NOL=1, Gene Adh, maxmer with bias correction



**Figure 16** - VMI 3D heat map – *Drosophila ananassae* 12 sequences, L=860 nucleotides, NOL=1, Gene Adh, maxmer with bias correction - this plot is generated from the same 2D heat map data

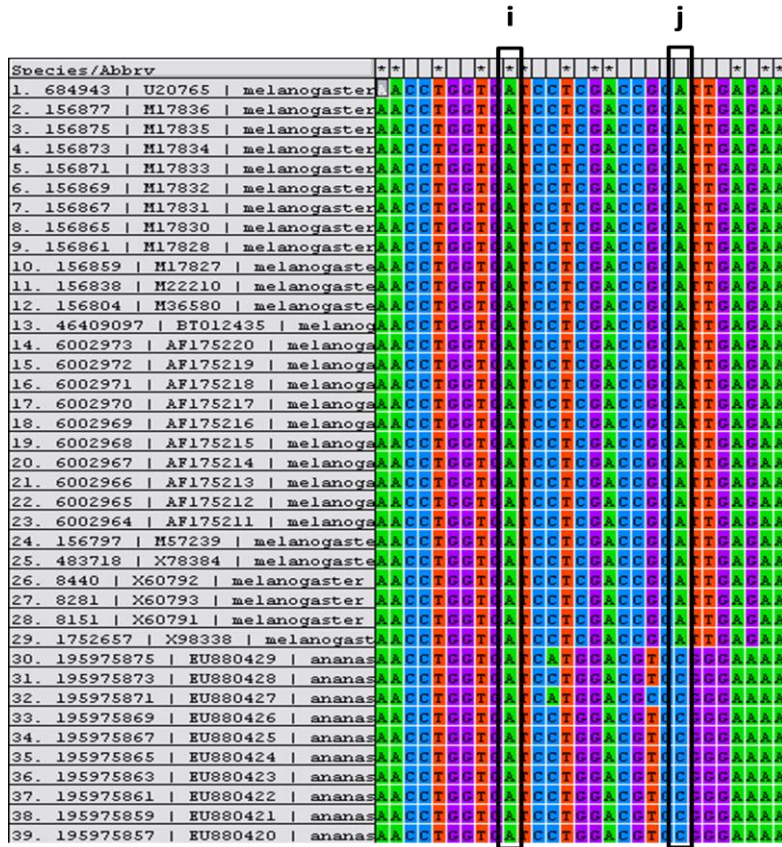
### 14.3. How to calculate Vertical Mutual Information

To obtain VMI first we must calculate nucleotide frequencies for position pairs (*i*, *j*) scanning all *i* sites versus all *j* sites, where *j* > *i*.

**VMI** is a functional **VMI(*i*, *j*)** and is represented by a heat map, given by,

$$VMI(i, j) = \sum_{m=\{A,G,T,C\}} \sum_{n=\{A,G,T,C\}} p_{mn}(i, j) * \log \frac{p_{mn}(i, j)}{p_m(i)p_n(j)} \quad (6)$$

where *i* and *j* are distinct positions and *m* and *n* are possible nucleotides  $\epsilon \{A, G, T, C\}$ .



**Figure 17** - Calculus of VMI cross-correlation – obtaining vertical  $p_{mn}(i, j)$ ,  $p_m(i, j)$  and  $p_n(i, j)$ , for  $j > i$ , and  $i, j = [1, L]$

The marginal probabilities ( $p_m(i, j)$  and  $p_n(i, j)$ ) are calculated by,

$$p_m(i, j) = \sum_{n=\{A,G,T,C\}} p_{mn}(i, j) \quad (4)$$

and

$$p_n(i, j) = \sum_{m=\{A,G,T,C\}} p_{mn}(i, j) \quad (5)$$

## 15. Horizontal Mutual Information (HMI)

Algorithm 07 (A7), tab HMI, is the HMI algorithm that calculates Horizontal Mutual Information. Once there are four possibilities:

- Mincut:
  - without bias correction
  - with bias correction
- Maxmer:
  - without bias correction
  - with bias correction

we decide to help the user with the button <calc all> (see below). That means, if you click <calc each> MIA will obey your choice of Mincut or Maxmer and observe if "bias correction" check box is on or off. Once each calculation may spent a lot of time (from a few minutes to many minutes depending on how large your sequence is, how many sequences you presented and the Number of Letters (NOL) chosen) we decided to create <calc all> and let MIA do all this job.

NCBI | Gbk-Fasta | Alignment | Purging | Consensus | VMI | **HMI** | JSD | Cluster | Entropy | Root

**HMI - Horizontal Mutual Information**

Organism: Drosophila - Adh -

Gene: Adh

Num of letters: 2

Frame: 0

Offset: 0

min max calc each calc all

Gene CDS Exon

☐ Save data ☒ See image ☐ Save image ☐ Recalculate ☒ Normalized

☐ bias correction ☒ mnat

Image: png DPI: 300

Input: Drosophila\_(max/min)\_Gene\_Adh\_100L\_cutoff7\_consensus.fasta

Output: hmi\_Drosophila\_maxmer\_(species)\_Gene\_frame0\_NOL2\_100L\_cutoff7\_mij.txt

Message:

Horizontal Mutual Information | Save | Exit | Clear

**Figure 18** – Fifth Tab, Consensus, replace gaps by vertical consensus nucleotide

#### Options:

- "organism": the desired organism, don't change the string.
- "min/max": choose between mincut or maxmer sequences
- "calc each"/"calc all": calc each chosen screen parameters, or calc all four options ("min and max" versus "wo/with bias correction").
- "gene": the desired gene, don't change the string.
- "num of letter": you can vary NOL from 1 to any other integer (recommended < 10).
- "frame": in this version maintain equal 0 (zero). All three frames.
- "offset": the initial horizontal displacement, recommended maintain 0 (zero)
- "Gene/CDS/Exon": recommended Gene (see your data).

#### Other options:

- ""Save Data": saves "summary" file at the end of all calculations. Summary file is a summary in a table format plus the ANOVA result.



Once you end the whole job we recommended turn this flag off, and if possible do a backup.

- "See image": displays the resulting image (a distribution, a histogram, etc). Be careful that there are so many images equal to the number of selected options (species versus min/max versus bias correction). That can be dozen of images that you must spend a time analyzing and closing. If you close MIA, all images will be closed too. Better save all without seeing and the choose one or two to see or open them in the operational system.
- "Save image": if checked saves image in its directory (e.g., ~/data/Drosophila/image) and you can choose which type (png, tif and jpg) and the resolution (in dpi, we recommended 120 for draft, 300 for a good resolution and 600 for a very good resolution).
- "Recalculate": once calculated, MIA will not do the job again, unless you turn this check box on.
- "Normalized": if checked divides all Mutual Information values by the number of letters, normalizing them. Let this button initially off to observe how MI varies. If you turn off and after turn on, MIA will not recalculate all values again, is a fast operation.
- "bias correction": apply Roulston bias correction

- "mnat": once most of the calculation results in values less than 1 Nat, this option multiplies each value by 1000 given in mnat (millinat) unit.

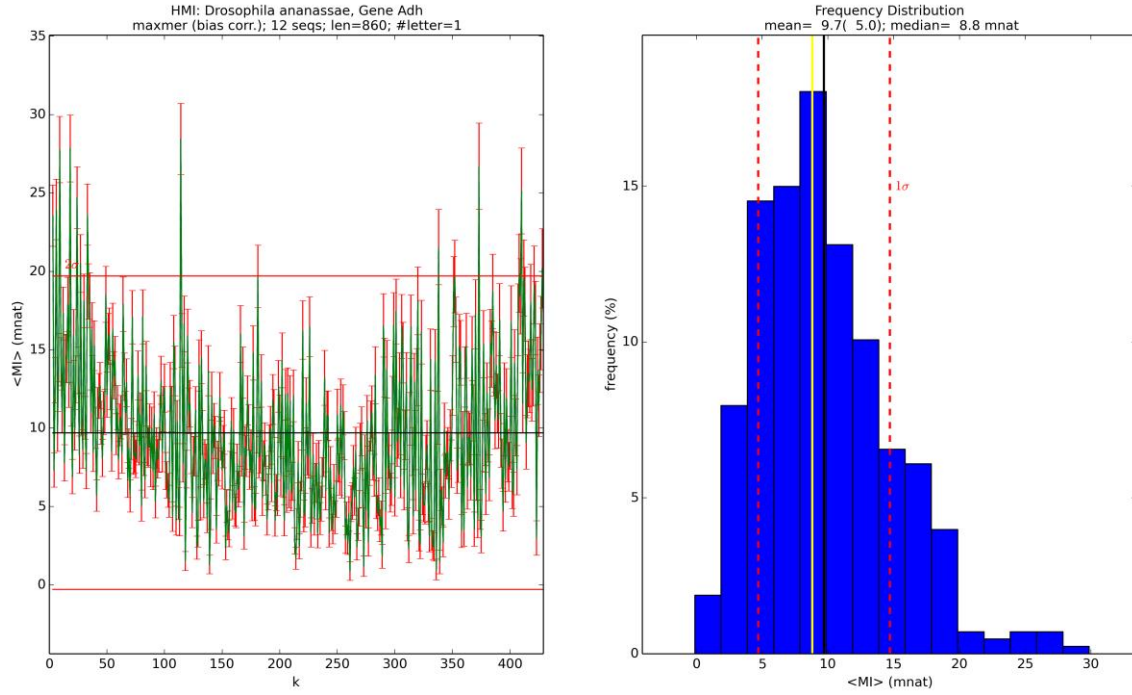
Run HMI:

- click on <Horizontal Mutual Information> to calculate and display HMI

Input / Output files:

- Input: 2 fasta files per species:
  - "Drosophila \_maxmer\_Gene\_Adh\_100L\_cutoff7\_consensus.fasta"
  - "Drosophila \_mincut\_Gene\_Adh\_100L\_cutoff7\_consensus.fasta"
- Output: many files in dictionary python format:
  - (mincut x maxmer) x (wo correction, with correction)
  - HMI\_Drosophila\_min/max\_(species)\_Gene\_(gene)\_frameX\_NOL(x)\_100L\_cutoff7\_(type)(bias).txt
  - where:
    - gene: like ADH, AMY, etc.
    - frameX: choose frame equal 0 (all frames, but MIA can calculate frame 1, 2 or 3.
    - NOL(x) = number of letters: word size (e.g. from 1 to 7)
    - type:
      - \_mi: mutual information (MI)
      - \_se: Standard Error from MI
    - bias: "\_bias\_corr" for bias correction or nothing without bias correction.
  - VMI\_params\_Drosophila\_min/max\_Gene\_(gene)\_frameX\_NOL(x)\_100L\_cutoff7\_(bias).txt
    - "params" is a summary table with all information.

## 15.1. Horizontal Mutual Information



**Figure 19** – At the left the HMI graphic for *Drosophila ananassae* 12 sequences,  $L=860$  nuc,  $NOL=1$ , Gene Adh, maxmer with bias correction - the horizontal red line stands for 2 standard deviation of the distribution. At the right we see the frequency distribution graphic with 4 vertical lines: two red lines for standard deviation, in yellow the median and in black the mean of the VMI distribution.

## 15.2. How to calculate Horizontal Mutual Information

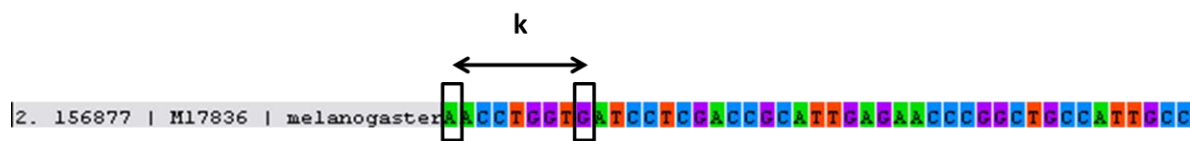
**HMI** is defined as an autocorrelation measure between two pair of positions  $k$  units of distance away from each other (for  $k \in [1, L/2]$ , where  $L$  is the length from one sequence). Here, “ $k$ ” represents the distance between two residues, its interval ranges from 3 to  $L/2$ , with step equal to one (all three frames are calculated). Another parameter to this definition is the length of the word. In biology DNA, RNA and Protein interactions occur with motifs that usually have little interaction length with  $L_{int} \sim < 10$ . In the present work we calculated **HMI** only for words with length equal 1 (one letter).

Fixing a value for ‘ $k$ ’, all sequence is scanned counting nucleotide pairs,  $(m,n) \in \{AA, AG, \dots CC\}$ . Here,  $p_{mn}$  represents the probability of each possible  $(m,n)$  pairs, where  $m \in \{A,G,T,C\}$  and  $n \in \{AGTC\}$ . Thus HMI is given by,

$$HMI(k) = \sum_{m \in \{A,G,T,C\}} \sum_{n \in \{A,G,T,C\}} p_{mn}(k) * \log \frac{p_{mn}(k)}{p_m(k)p_n(k)} \quad (3)$$



The marginal probabilities are explained in VMI (see above).



**Figure 20** - How to calculate HMI autocorrelation – obtaining  $p_{mn}(k)$  for  $k$  distance where  $m, n \in [A, G, T, C]$

## 16. Jensen-Shannon Distance

Algorithm 08 (A8), tab JSD, is the algorithm that calculates the Jensen-Shannon Divergence (or distance). JSD is a distance functional derived from Kullback-Leibler Divergence because this last is not symmetry and distance definitions determines  $d(A, B) = d(B, A)$ . It measures the distance between two distributions and is applied in MIA to discriminate species. In other words to measure the distance between species distribution pair wises.

**Figure 21** - JSD algorithm result in a table and a histogram with all distance between species distributions

#### Options:

- "organism": the desired organism, don't change the string.
- "gene": the desired gene, don't change the string.
- "min/max": choose between mincut or maxmer sequences
- "calc each"/"calc all": calc each chosen screen parameters, or calc all four options ("min and max" versus "wo/with bias correction").
- "num of letter": you can vary NOL from 1 to any other integer (recommended < 10).
- "Vertical Entropy/VMI/HMI": choose which method you want
- "frame": in this version maintain equal 0 (zero). All three frames. But only for HMI, VMI does not have frame.

#### Other options:

- "Save Data": saves "summary" file in the end of all calculations. Summary file is a summary table like plus the ANOVA result. Once you did the whole job we recommended turn this flag off, and if possible do a backup.
- "See image": displays the resulting image (a distribution, a histogram, etc). Be careful that there are so many images equal to the number of selected options (species versus min/max versus bias correction). That can be dozen of images that you must spend a time analyzing and closing. If you close MIA, all images will be closed too. Better save all without seeing and then choose one or two to see or open them in the operational system.
- "Save image": if checked saves image in its directory (e.g., ~/data/Drosophila/image) and you can choose which type (png, tif and jpg) and the resolution (in dpi, we recommended 120 for draft, 300 for a good resolution and 600 for a very good resolution).
- "Recalculate": once calculated, MIA will not do the job again, unless you turn this check box on.
- "bias correction": apply Roulston bias correction
- "mnat": once most of the calculation results in values less than 1 Nat, this option multiplies each value by 1000 given in mnat (millinat) unit.

#### Run JSD:

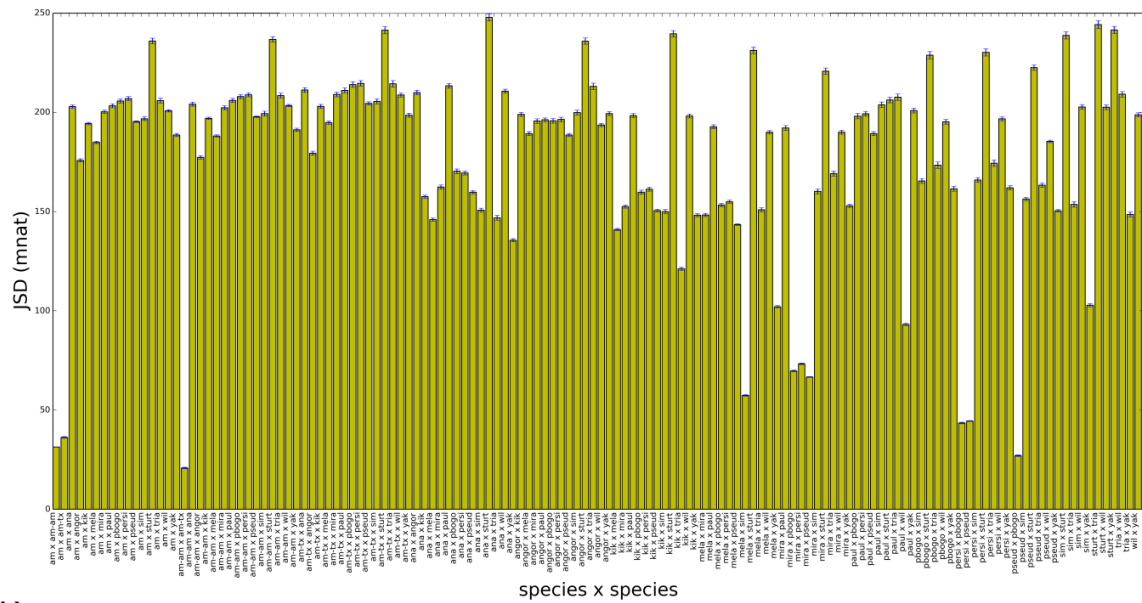
- click on <Jensen-Shannon Divergence> to calculate it.

Input / Output files:

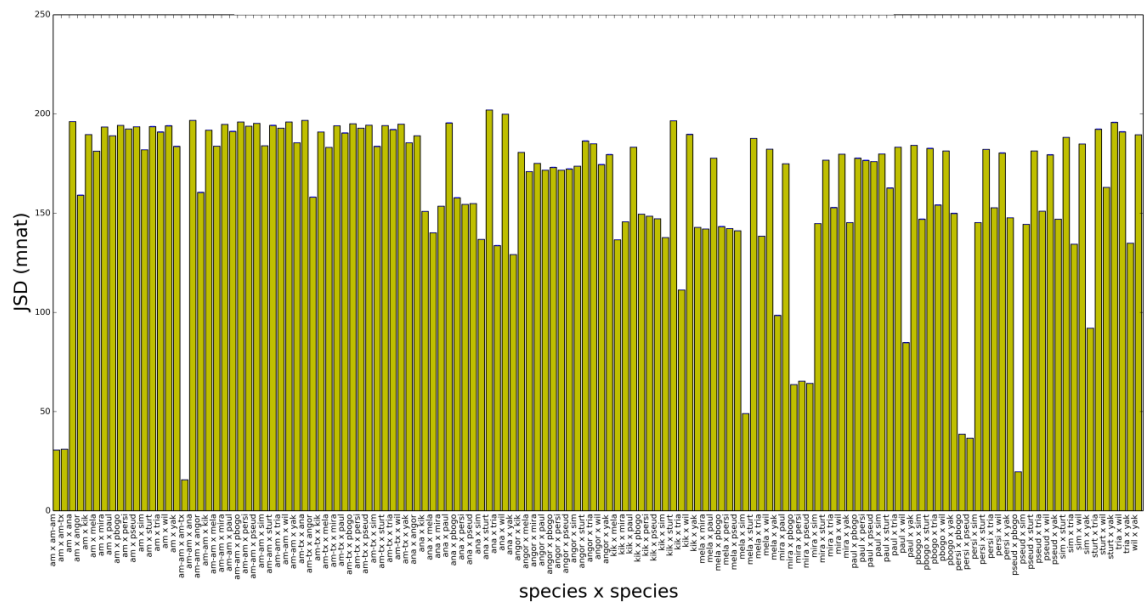
- Input:
  - VMI or HMI output
- Output:
  - VMI: JSD\_VMI\_Drosophila\_min/max\_Gene\_(gene)\_NOL(x)\_100L\_cutoff7\_(bias).txt
  - VH: JSD\_VSH\_Drosophila\_min/max\_Gene\_(gene)\_NOL(x)\_100L\_cutoff7\_(bias).txt
  - HMI: JSD\_HMI\_Drosophila\_min/max\_Gene\_(gene)\_frameX\_NOL(x)\_100L\_cutoff7\_(bias).txt
  - where:
    - gene: like ADH, AMY, etc.
    - frameX: choose frame equal 0, only for HMI
    - NOL(x) = number of letters: word size (e.g. from 1 to 7)
    - bias: "\_bias\_corr" for bias correction or nothing without bias correction.
  - ASCII tables:
    - The same name as above for JSD values.
    - same name + '\_se.txt' for SE(JSD) values.
    - same name + '\_summay.txt' summarizing all JSD between species distributions.

## 16.1. JSD from Vertical Entropy

a)



b)



**Figure 22a** - JSD histogram for species pairwise HMI distributions from maxmer sequences with NOL=1 and 2SE, a) with bias correction and b) without: "am" for *D. americana*, "am-am" for *D. americana americana*, "am-tx" for *D. americana texana*, "ana" for *D. ananassae*, "angor " for *D. angor*, "kik" for *D. kikkawai*, "mela" for *D. melanogaster*, "mira" for *D. miranda*, "paul" for *D. paulistorum*, "pbogo" for *D. pseudoobscura bogotana*, "persi" for *D. persimilis*, "pseud" for *D. pseudoobscura*, "sim" for *D. simulans*, "sturt" for *D. sturtevantii*, "wil" for *D. willistoni*, "yak" for *D. yakuba*.

## 16.2. Jensen-Shannon Distance Definition

Jensen-Shannon Divergence is a method of measuring distances between two or more distributions. Once we are interesting in the distance between each two Mutual Information distribution we present JSD as  $JSD(P || Q)$ . Here P and Q are HMI or VMI distributions. Thus, JSD is given by

$$JSD(P || Q) = H\left(\frac{P+Q}{2}\right) - \frac{1}{2} (H(P) + H(Q)) \quad (10)$$

## 16.3. JSD Standard Error - SE(JSD)

Besides the calculation of JSD for all species distribution pairwises, we also attempted to calculate the SE. Once Mutual Information and Jensen-Shannon Distribution are not linear functions, we must propagate it empirically. That means, in the MI Space we must calculate all four possible distances crossing species(p) versus species(q). Thus, we calculated  $(D_p+SE \text{ and } D_p-SE) \times (D_q+SE \text{ and } D_q-SE)$ , where p and q are different species indexes and D is VH, VMI or HMI distributions. Then we applied JSD to all four possible combined distributions and evaluate  $\max(JSD)$  and  $\min(JSD)$ . The empirical SE is defined as:

$$SE[JSD(P||Q)] = \left( \max(JSD[\text{any } D_p+-SE||D_q+-SE]) - \min(JSD[\text{any } D_p+-SE||D_q+-SE]) \right) / 2 \quad (11)$$

## 17. Hierarchical Cluster

Algorithm 09 (A9), tab Cluster, is the algorithm that calculates clusters from distance matrices. Here Hierarchical Cluster is displayed as Dendrograms only with the intuit of visualization.

The screenshot shows a software window with multiple tabs at the top: NCBI, Gbk-Fasta, Alignment, Purging, Consensus, VMI, HMI, JSD, Cluster, Entropy, and Root. The 'Cluster' tab is active. Below the tabs, the 'Hierarchical Cluster' section contains the following controls:

- Organism: Drosophila (selected from a dropdown)
- Gene: Adh (text input)
- Buttons: min, max, calc each, calc all
- Num of letters: 2 (text input)
- Buttons: Gene, CDS, Exon
- Analysis: Vertical Entropy, Vertical MI, Horizontal MI (radio buttons)
- Frame: 0 (text input)
- Method: Complete, Single, WPGMA, Centroid (radio buttons)
- Leaf threshold: 0.025 (text input)
- Buttons: ☒ See image, ☐ Save image
- Buttons: ☐ bias correction, ☒ mnat
- Image: png (dropdown), DPI: 300 (text input)
- Input: JSD\_VSH\_Drosophila\_maxmer\_Gene\_Adh\_NOL2\_100L\_cutoff7.txt (text input)
- Message: (empty text input)

At the bottom of the window, there is a large light blue rectangular area and a row of buttons: Hierarchical Cluster, Save, Exit, and Clear.

Figure 23 - Hierarchical cluster tab

Options:

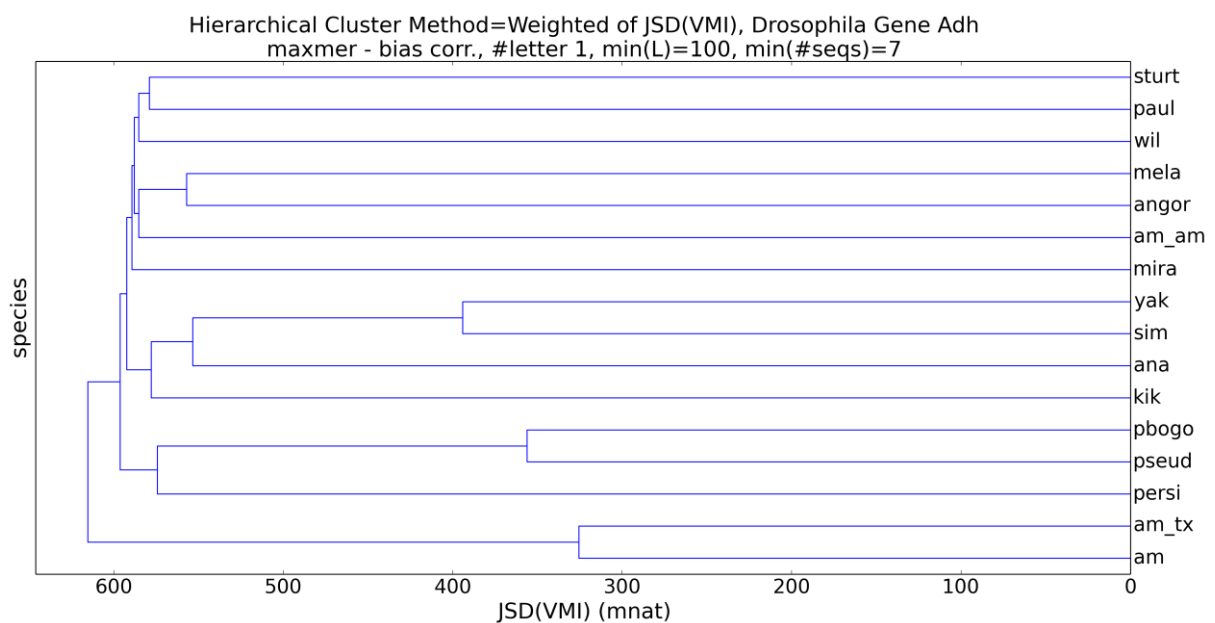
- is almost the same as the last tabs
- "Method": there are three possible methods:
  - Complete (maximum distance)
  - Single (minimum distance)
  - WPGMA (weighted pair group method with averaging) - recommended
  - Centroid

- "Leave threshold": is a distance defined by the user to group tips with a same color

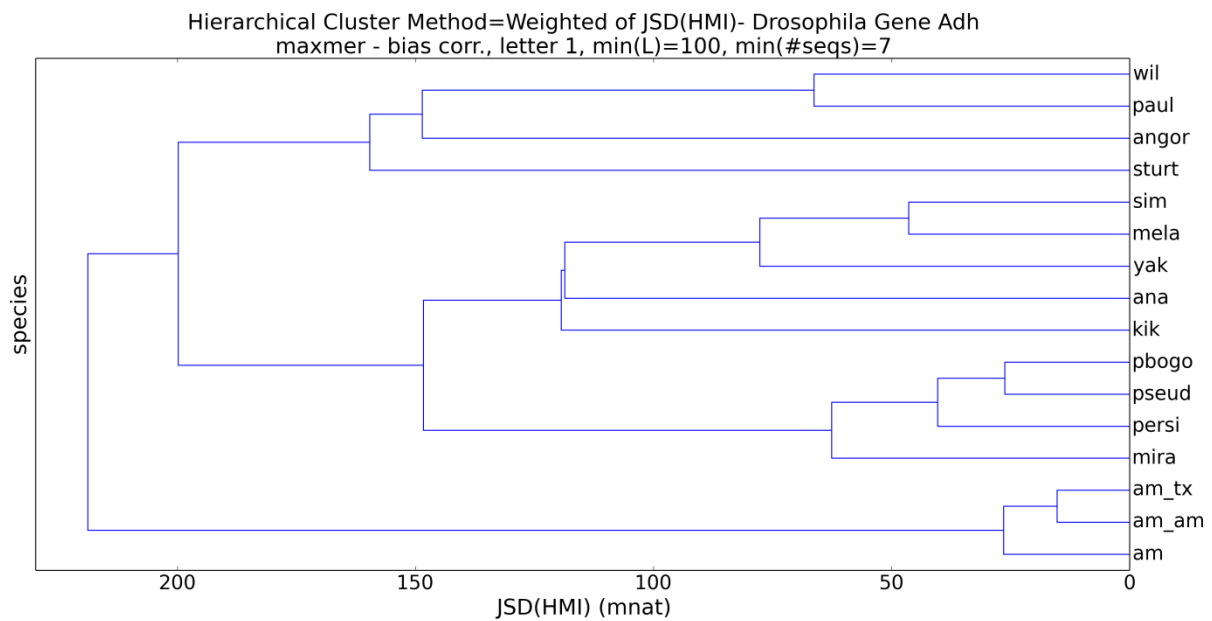
Run HC:

- click on <Hierarchical Cluster> to display the dendrogram.
- Input:
  - JSD output file
- Output:
  - Dendrogram images

## 17.1. Hierarchical Cluster Dendrogram



**Figure 24** - Hierarchical cluster dendrogram from VMI, method=WPGMA, Gene Adh, maxmer with bias correction, NOL=1



**Figure 25** - Hierarchical cluster dendrogram from HMI, method=WPGMA, Gene Adh, maxmer with bias correction, NOL=1

## 18. Shannon Entropy (simulation)

Algorithm 10 (A10), tab Entropy, is the Simulation of Shannon Entropy.

NCBI Gbk-Fasta Alignment Purging Consensus VMI HMI JSD Cluster Entropy Root

**Shannon Entropy**

Database: Nucleotide

Num of letters: 2

Num of experiments: 10

Start at: 5

Length simulation: 500

Image: png DPI: 300

☐ Save data ☒ See image ☐ Save image

Output: shannon\_random\_DNA\_LetterNNN\_ExpNNN\_dic.txt

Message:

Shannon Entropy Save Exit Clear

**Figure 26** - Shannon Entropy Simulation



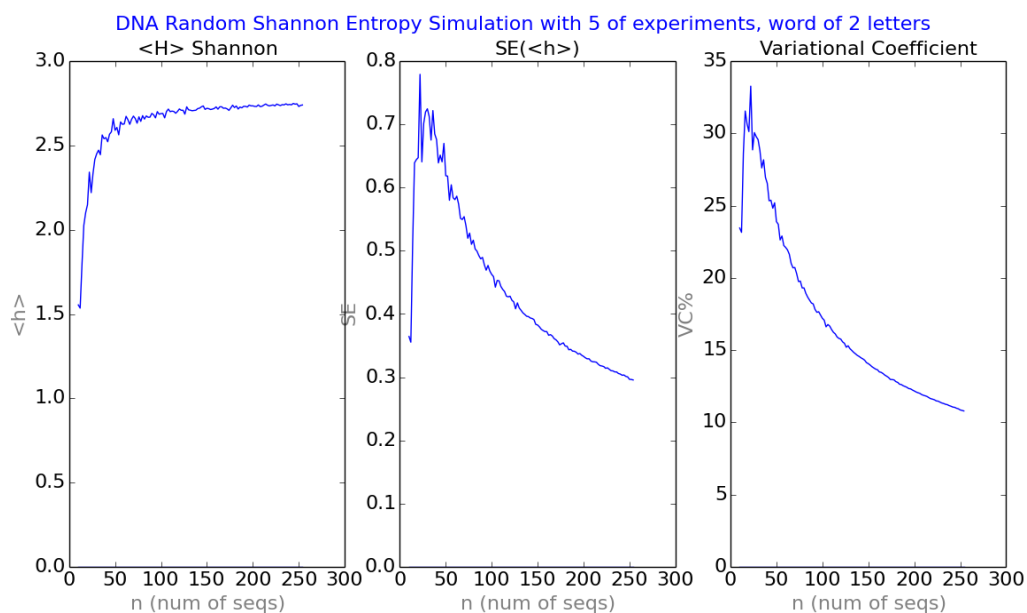
#### Options:

- "Database": choose between Nucleotide (4 types = {A,G,T,C}, MER=2 in bits) or Protein (20 types, MER=4.322 in bits)
- "Number of letters": choose between 1 and any integer number. As NOL increases MER increase and the simulation will spent more time.
- "Number of experiments": choose between and any integer number. Recommended a number near 25. If Number of experiments is less than 10 we observe a noise in the simulation, as it increases the noise tends to zero.
- "Start": never let start less than 3. With shorter string to simulate there is instability in the simulation. Recommended 5.
- Length of simulation: any number greater than 50. As NOL increases the length of simulation should also increase to analyze the upper limit.
- Other options: see previous tabs.

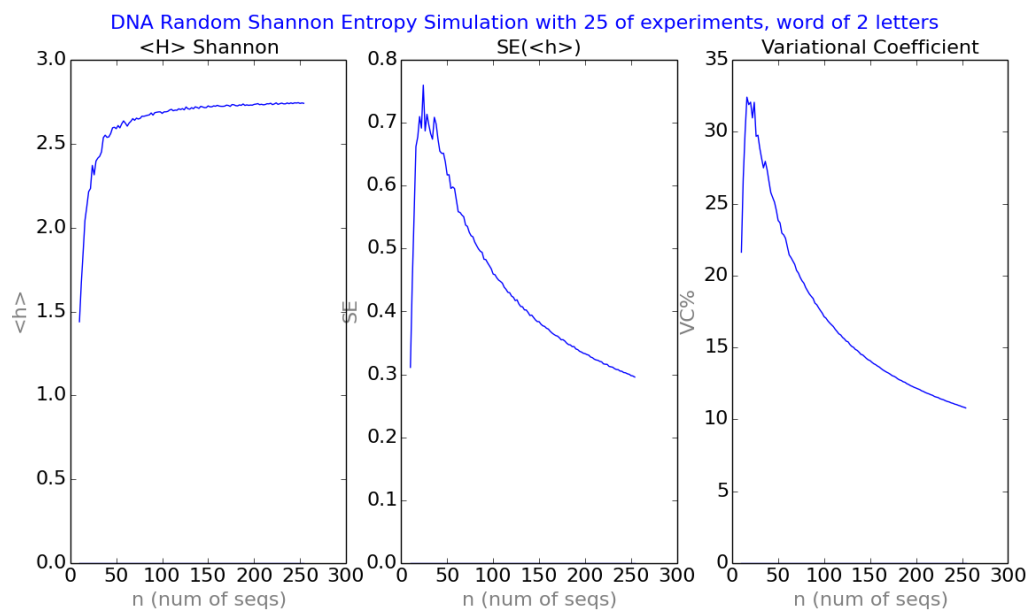
#### Run Shannon Entropy:

- click on < Shannon Entropy> to simulate and display graphics.
- Input:
  - None
- Output:
  - An ASCII table in python dictionary format with name like:  
  
shannon\_random\_(DNA/Nucleotide)\_LetterNNN\_ExpNNN\_dic.txt
  - where:
    - LetterNNN = NNN is the number of letters in the experiment
    - ExpNNN = NNN is the number of repeated experiments.

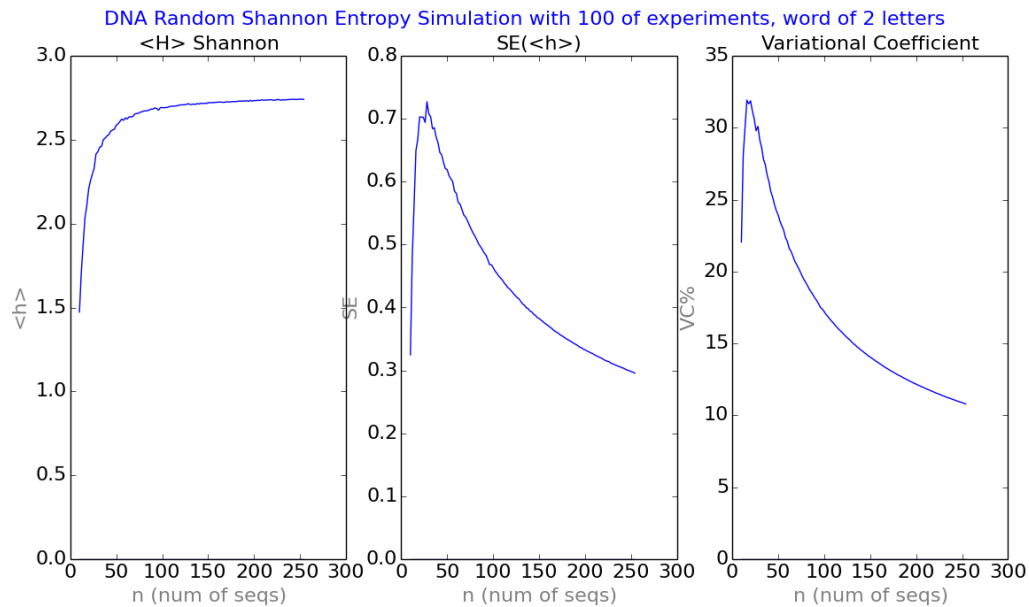
## 18.1. Shannon Entropy Simulation graphics



**Figure 27** - Shannon Entropy Simulation, word of 2 letters, 5 repeated experiment results in high noise



**Figure 28** - Shannon Entropy Simulation, word of 2 letters, 25 repeated experiment results in medium noise



**Figure 29** - Shannon Entropy Simulation, word of 2 letters, 100 repeated experiment results in low noise

## 19. Bibliography

Biopython community (2012). Biopython.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol. Biology Evol.* 27, 221–224.

Johnson, A.D. (2010). An extended IUPAC nomenclature code for polymorphic nucleic acids. *Genet. Popul. Anal.* 26, 1386–1389.

Roulston, M.S. (1999). Estimating the errors on measured entropy and mutual information. *Phys. D* 125, 285–294.

Flavio Lichtenstein

e-mail: flalix@gmail.com

Federal University of Sao Paulo

DIS Bioinformatics

Rua Pedro de Toledo, 669, 4º andar, fundos.

CEP 04039-032

Vila Clementino São Paulo - SP - Brasil