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

Sumário

1. Introduction

Mutual Information Analyzer (MIA) is a pipeline written in Python (Python Community, 2012) with the intent to retrieve, manipulate molecular sequences, calculate 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 allows one to discriminate species, and can be displayed as histograms or heat maps. 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.

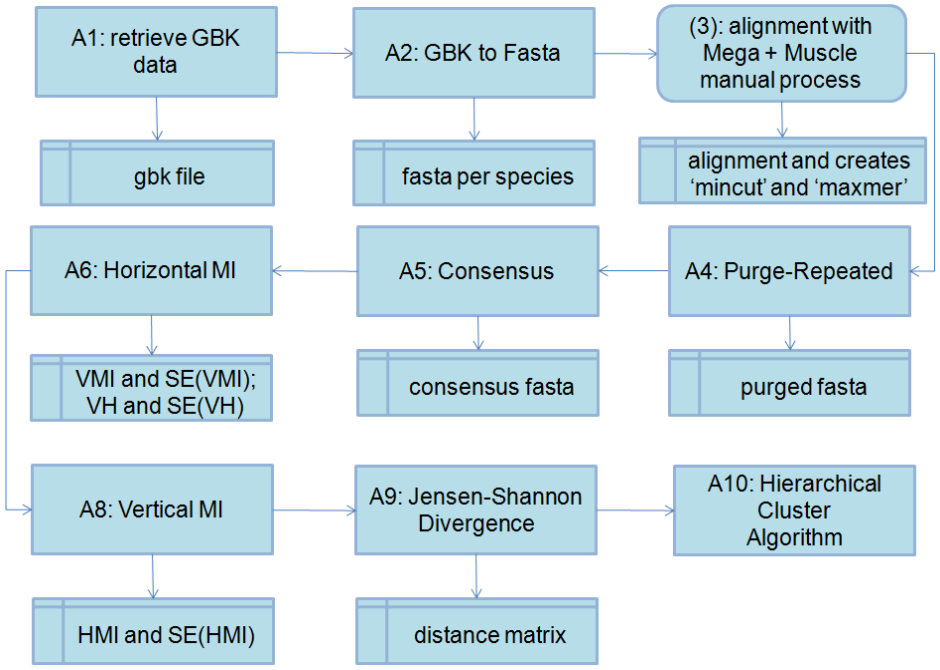

Figure 1 – Mutual Information Analyzer pipeline

At the end of the Alignment Algorithm MIA 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) can be calculated for "mincut" and "maxmer" and we also may apply a bias corrections (Roulston, 1999) for limited length sequences. Therefore, the gain or loss of information for "mincut" versus "maxmer", with or without bias correction, can be compared.

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

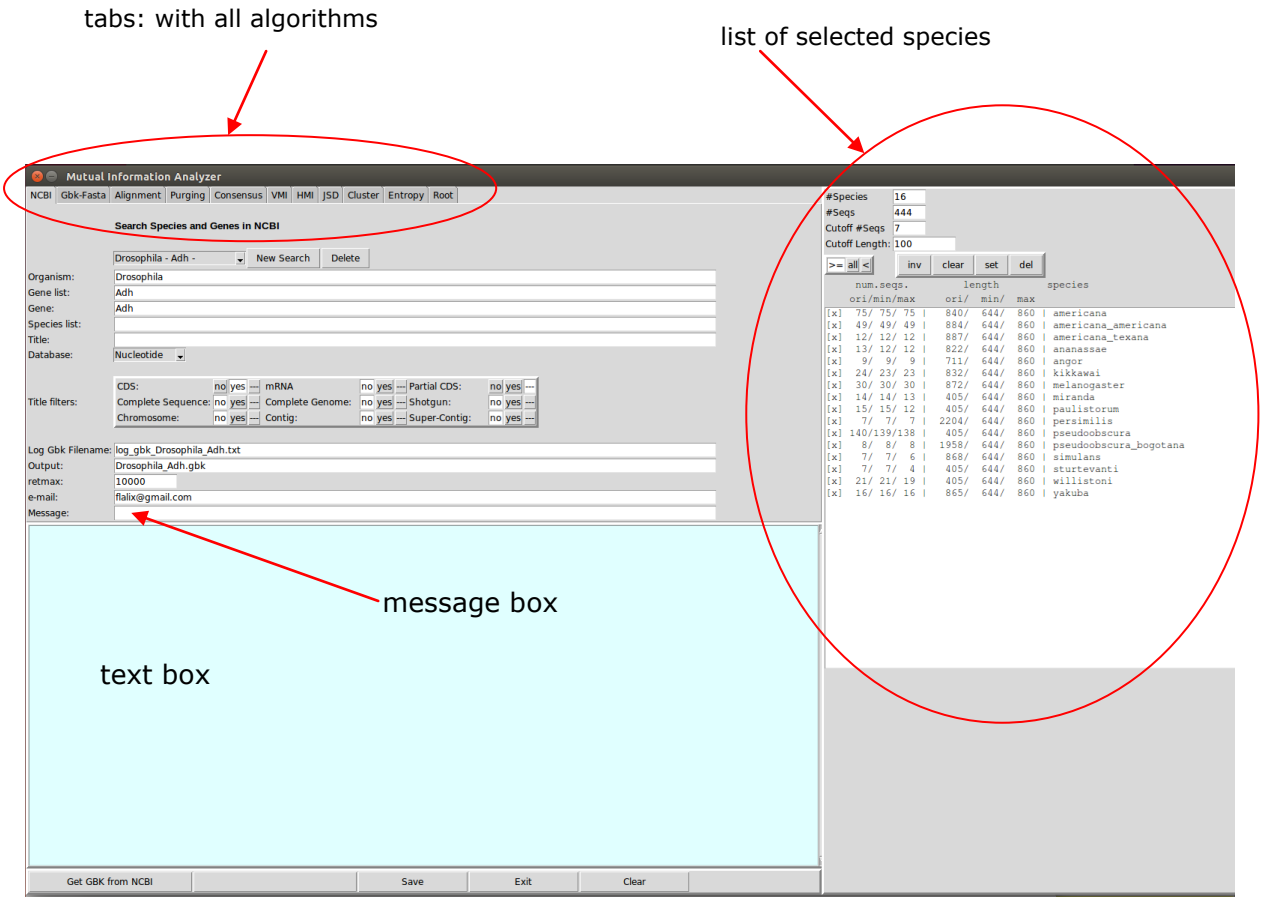

Figure 2 – General visualization of the first tab of MIA and Species List box at right

4. Species Listbox

At the end of the second algorithm Species Listbox will be filled with all species that could be 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 HMI with low SE. Therefore we should define a limit, e.g. at least 7 sequences. If 7 is chosen then many sequences some 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. Theses sequences must be ignored. If you are more exigent you can define the Cutoff Length 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 this analysis and all processed data will have those numbers in fasta file names or processed (calculated) file names.

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

The Species Listbox has:

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

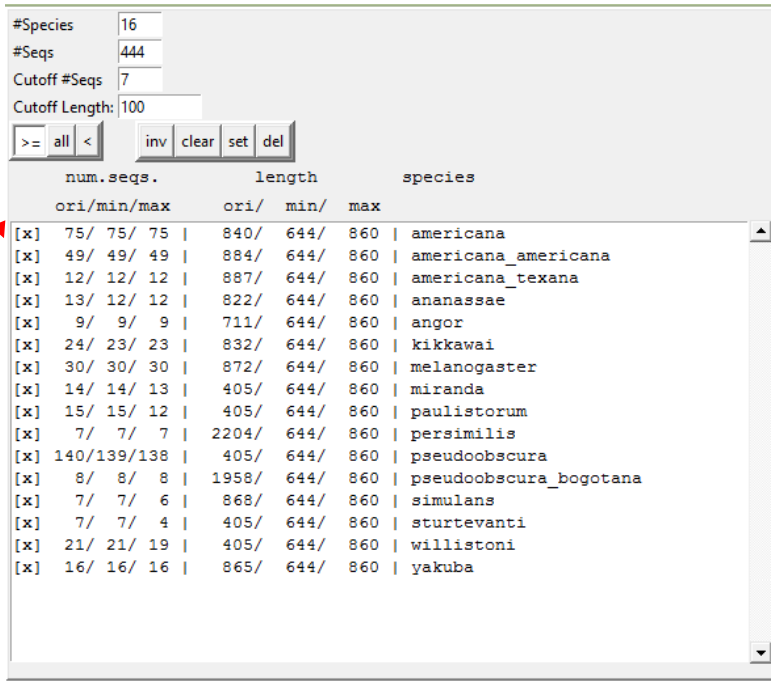

Figure 3 – Species List box and their command buttons

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, if it is installed in your computer. MIA also has two buttons one to see Mincut Sequences and other to Maxmer 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 analysiss explained bellow can be reproduced at NCBI site prompt (search).

The basic steps are:

- click in "New Search", then the screen is cleaned shortly thereafter
- define one Organism and one Gene / Title (better one Gene)
- some genes have many isoforms, therefore, in these cases, you may define a list of genes (a string, comma separated). This list is optative.
- MIA find all species for a determined Organism/Gene pair at NCBI, but if you know and want, 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 in GeneBank Title registers. As seen bellow, all filters (CDS, mRNA ... Super-Contig) also use Title to filter some conditions.

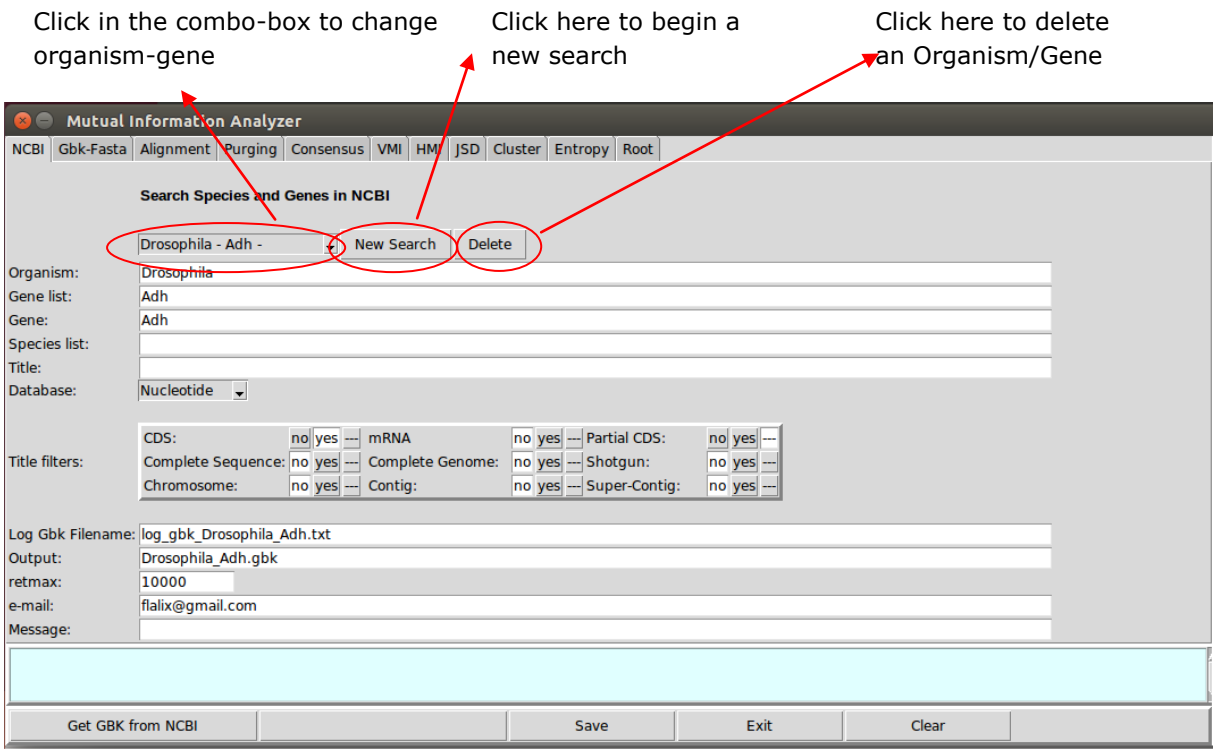

Figure 4 – MIA first tab, interfaces NCBI via a BioPython API.

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

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

It is obvious that there are many different analysiss:

a) only one gene
b) many genes
c) 2 kpb[1] 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)

In this version MIA is prepared to run, automatically, option a, "only one gene". But you may acquire the gene sequences many pair bases (pb) before the gene and many bases after it (option b), looking for interaction sites and conservative positions in the intergenic regions. This must be a manual operation and names must be MIA compliance.

The next step are the 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). This options are useful to study CDS and partial CDS without dealing with a whole genome, because this may have million to billion of base pairs.

"mRNA" should be avoided because sometime this sequences may come only with the exonic regions merged (cDNA) and in your analysis you are searching for Mutual Information and may 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 GeneBank 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 in these boxes.

[1] kbp - kilo base pair

- **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 / options defined.

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

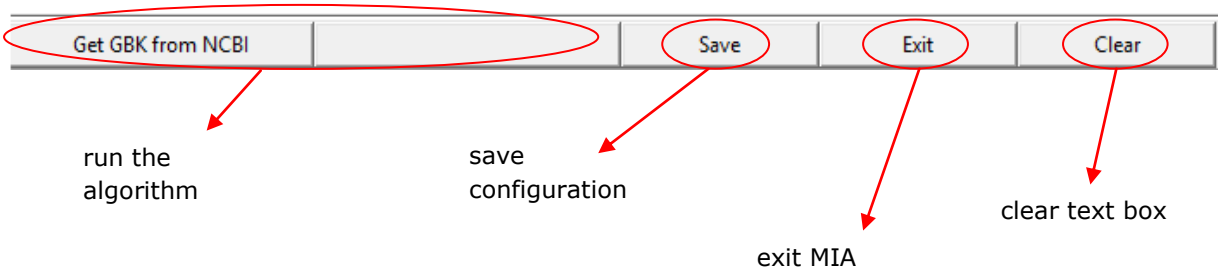

Figure 5 – At the bottom we find run, save, exit and clear command buttons.

Clear: it cleans the Text Box.

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

7. Directories / Roots

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

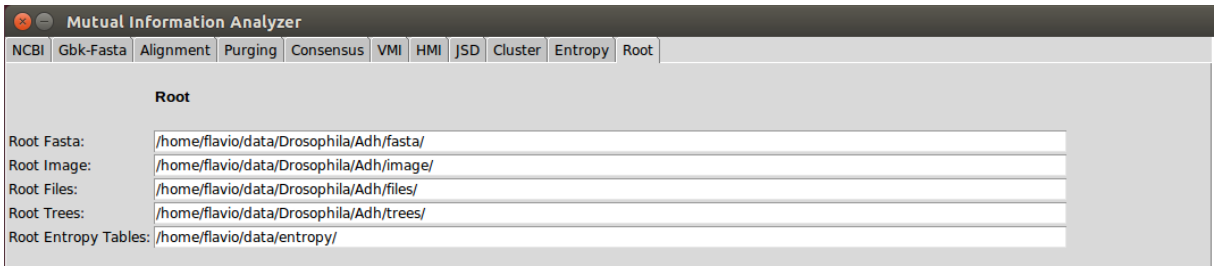

Figure 6 – The last tab shows the roots where fasta files, data tables, configs, logs and images can be found

- In Windows it looks like: "c:/users/Your_Name/data/ ...."
- In Linux it looks like: "home/Your_Name/data/ ...."

For each study (MIA analysis) you must define an Organism and a Gene. After saving this parameters click in <Save>, and MIA will create many subdirectories like:

- home/Your_Name/data/Your_Organism/Your_Gene/fasta for fasta files
- home/Your_Name/data/Your_Organism/Your_Gene/image for imagens (histograms, heatmaps, 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 in this simulations.

8. NCBI data acquisition

Algorithm 01 (A1), tab NCBI, is a data acquisition algorithm. It is a module that interacts with NCBI Nucleotide/Protein database retrieving GeneBank 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 with Nucleotides in its first version

and you may acquire Protein data, but you will not continue the analysis because it is not implemented or tested for proteins.

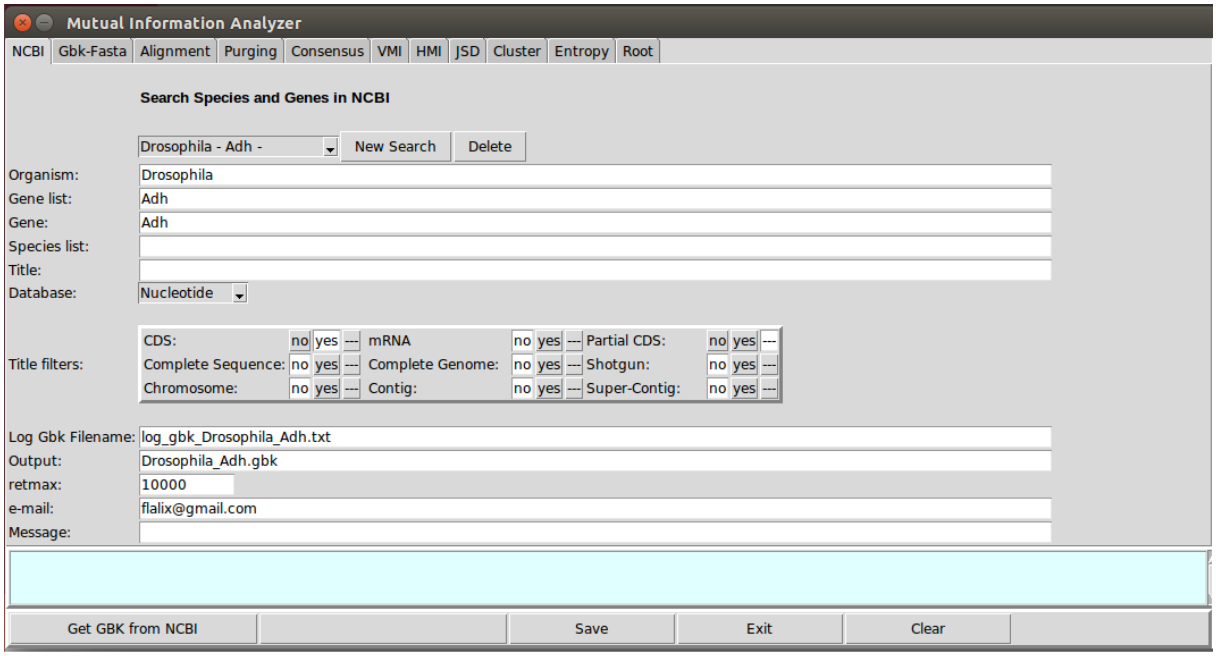

Figure 7 – How to get data from NCBI - in genebank format

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.

For each one 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 million to billion 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 file. Before starting this step, be careful, define "Cutoff Length", "Maximum Length Sequence" and "Minimum Length Sequence". The first of these parameter is used to define the output filename (see below) and must be maintained until the end of the analysis.

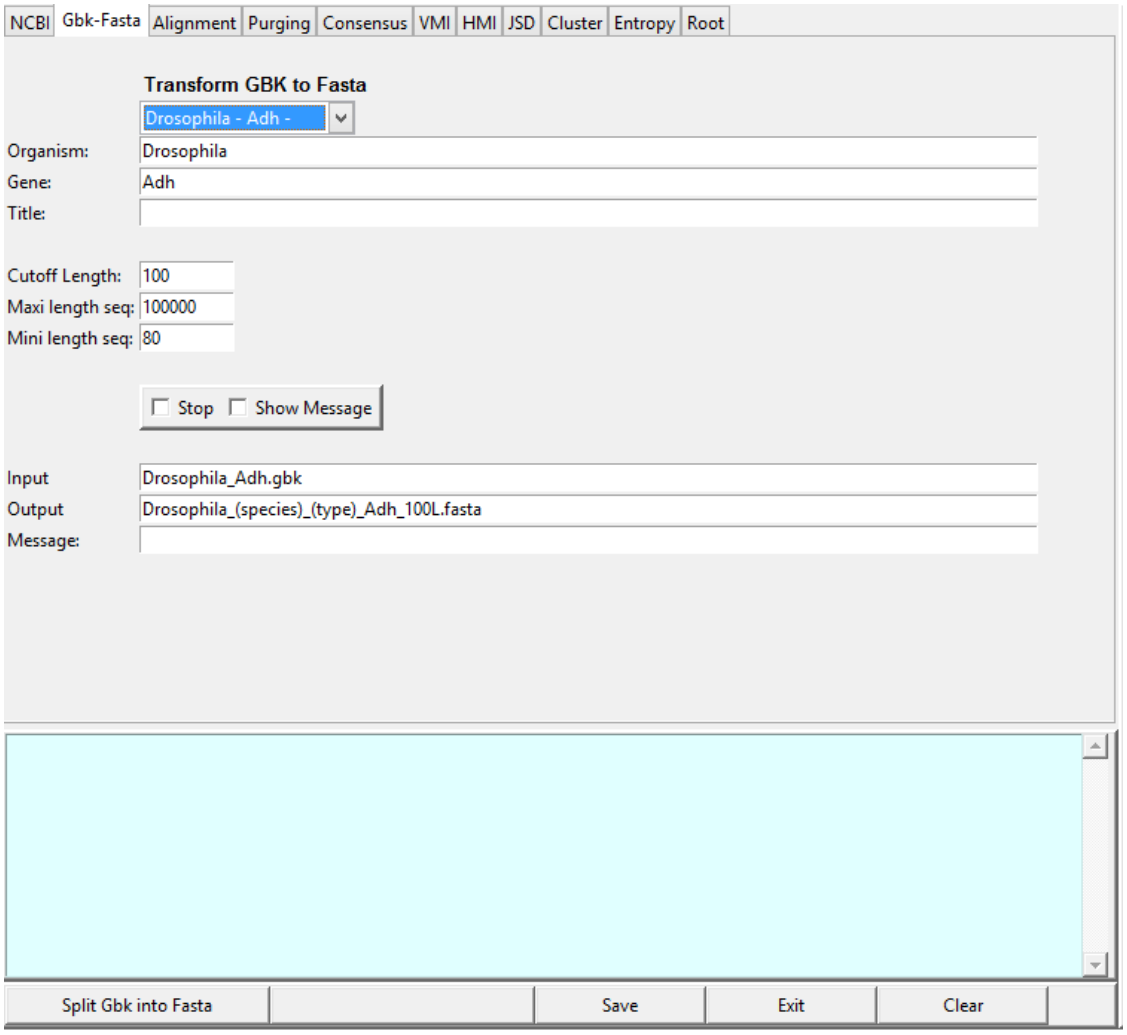

Figure 8 – Second Tab, Gbk-to-Fasta, splits gbk file in fasta files, one for each species before alignment

At the end of processing, the gbk-to-fasta file is sliced 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 "Drosophila_(species)_(type)_Adh.fasta"; one file per found species.

- Type is:
  - Gene, CDS, Exon, Intron or Protein
  - This algorithm creates five files for each species. But, as we observed, the most complete dataset is Gene and we suggest that the option "Gene" should 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, 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, do it manually or write to us.

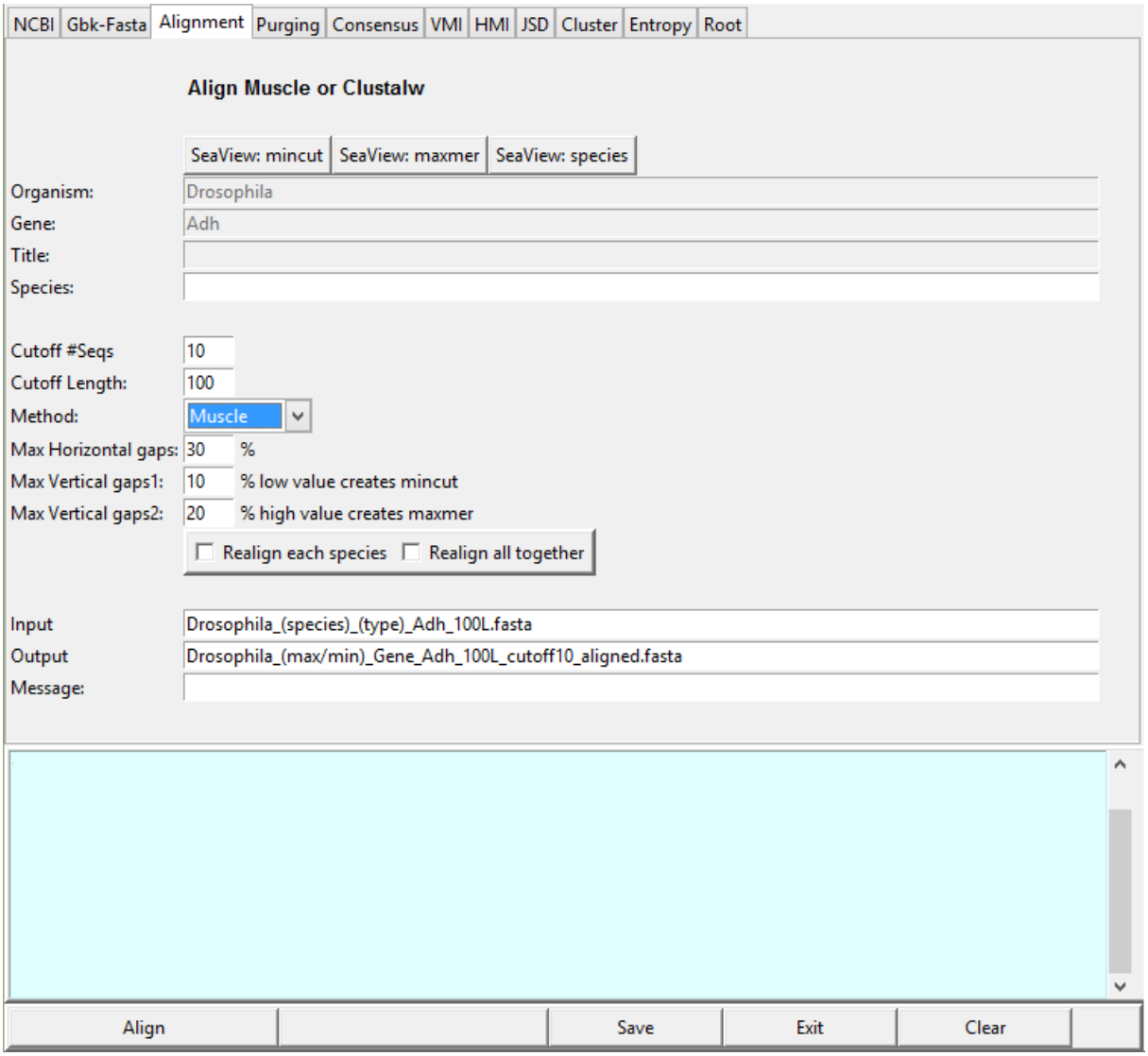

Figure 9 – Third Tab, Alignment, align each fasta file and at the end merge all them and realigned. All this alignments can be seen in Seaview. At the end of this algorithm "mincut" and "maxmer" fasta files are created.

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

Percentage filters, must defined:

- wished minimum percentage of sites occupied (without a gap) for a horizontal cutoff.

- wished minimum percentage of sites occupied vertically defines "maxmer" sequences,
- wished maximum percentage of sites occupied vertically defines "mincut" sequences.

Input / Output files:

- Input: fasta file name like "Drosophila _(species)_(type)_Adh.fasta"
- Output: 2 fasta files 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

11. How the algorithm works?

First MIA aligns each species sequences. Afterwards it merges all aligned sequences in a unique fasta file and aligned again.

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, or by cutting a bunch of sequences full of gaps or cutting vertical positions full of gaps. At the end, in another algorithm vertical positions must be 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". In this columns that exceed this quantity will be deleted. You must define to limits: Maximum Vertical Gaps 1 and 2. The first is a low value results in "mincut" fasta file, and the second with a higher value, allow more gaps, results in the "maxmer" fasta file.

In the other hand MIA also maintain sequences obeying the "maximum percentage of gaps". In this case a few columns with many gaps will be maintained resulting in the "mincut" fasta files.

Once in possession of "mincut" and "maxmer" fasta files MIA search for sequences that exceed the "maximum percentage of horizontal gapts". 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.

At the end, all sequences have the same length L_{min} for mincut fasta files and the L_{max} for maxmer fasta 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 to 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, 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 not be correct in terms of the reading frame. But, to Entropy and Mutual Information this important concept is not necessary in our analysis. Of course, is better a correct reading frame alignment and the conservation of all correct residues like in nature. But little changes will not result in drastic differences over Vertical Entropy, Vertical Mutual Information and Horizontal Mutual Information.

12. 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,

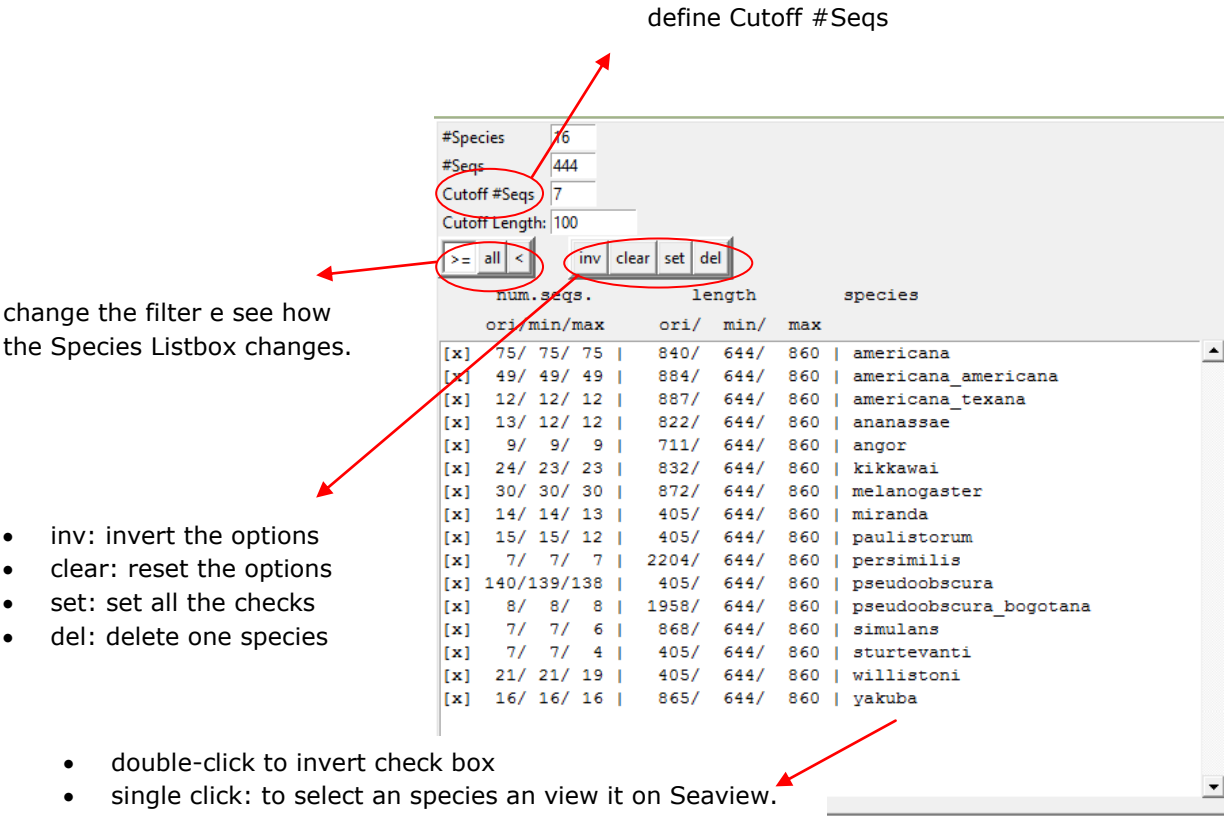

Figure 10 – Species List Box

13. 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 with the most consensus nucleotide.

It also looks for "strange" words in the description 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" tell 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 will echo more warning, leave it off.

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

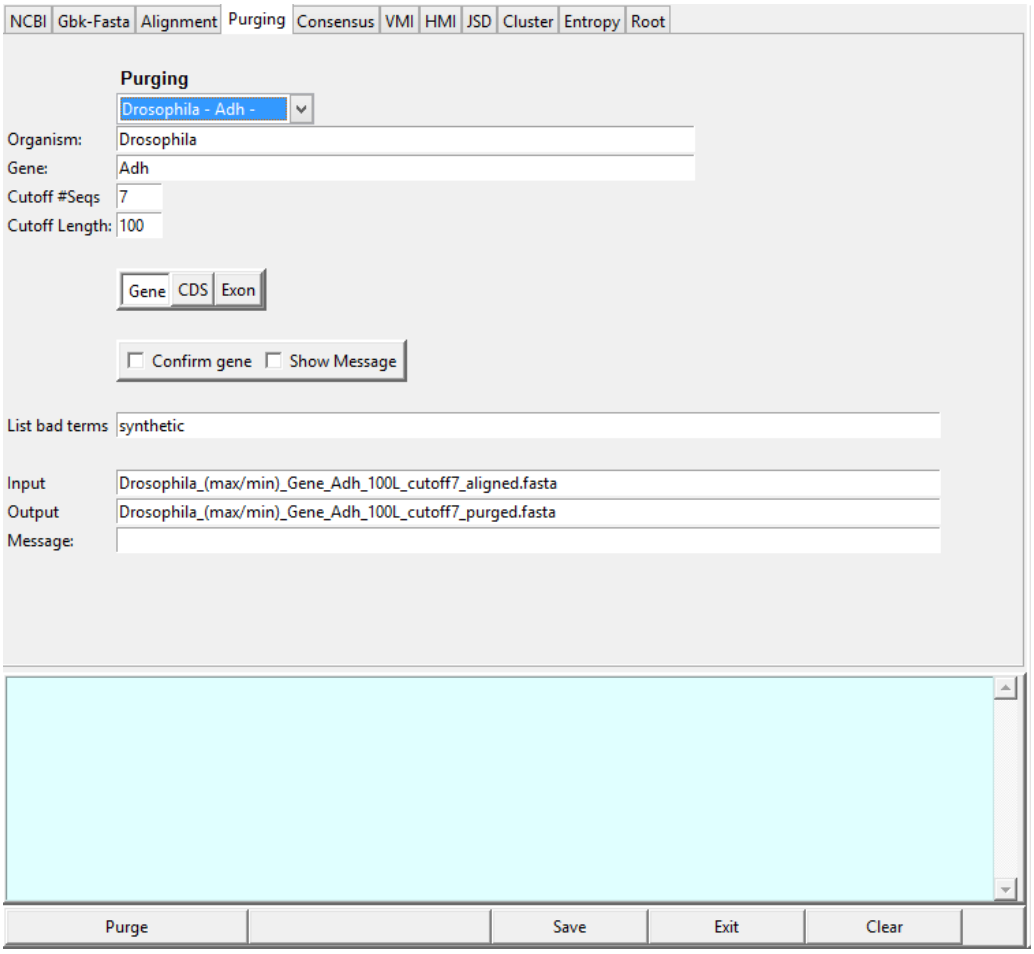

Figure 11 – Fourth Tab, Purging, search for ambiguous nucleotides and bad sequence description

- 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

14. Consensus

Algorithm 05 (A5), tab Consensus, is a gap replacer. It looks for position 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 replace all gaps with it. As we may notice, this action introduce information.

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

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

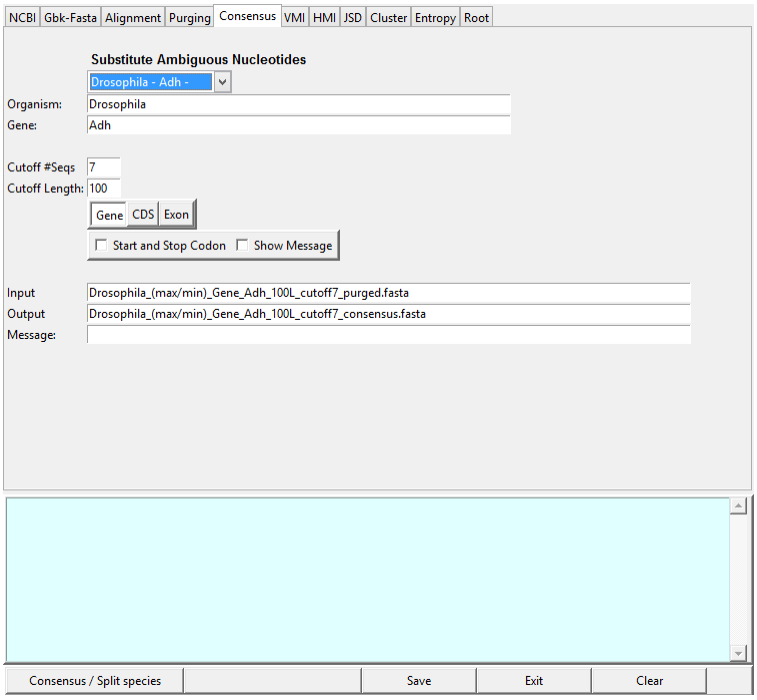

Figure 12 – Fifth Tab, Consensus, replace gaps by vertical consensus nucleotide

Input / Output files:

- Input: 2 fasta files per species:
    - "Drosophila _maxmer_Gene_Adh_100L_cutoff7_purged.fasta"

[2] For nuclear eukaryotes. For mitochondria and not prokaryotes these Stop Codons are not correct.

- "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"

15. 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
- Mincut with bias correction
- Maxmer without bias correction
- Maxmer with bias correction

we decide to help the user with the button <calc all>. 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 this calculations may spent a lot of time (from a few minutes to a few hours 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.

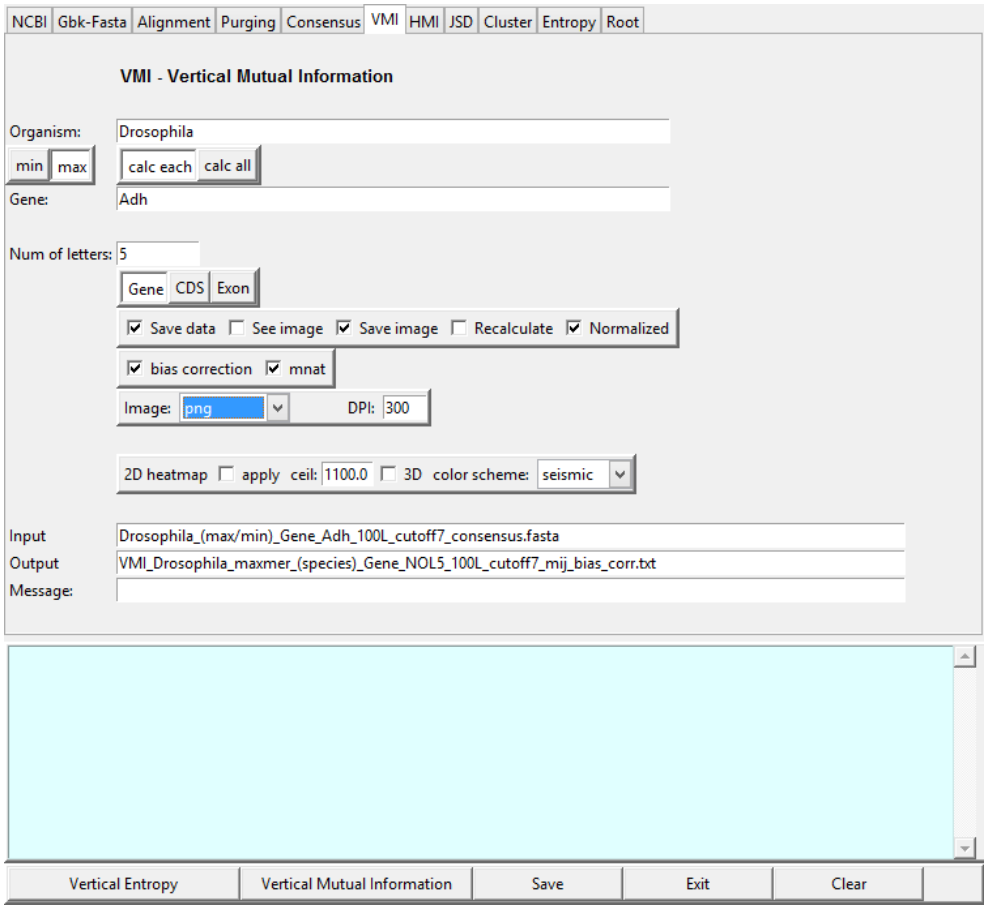

Figure 13 – Sixth Tab, VMI, calculates VH and VMI and display their graphics

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": calc each chosen screen parameters, or calc all four options ("min and max" versus "without/with bias correction").
- "num 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 in the end of all calculations. Summary file is a summary of all data as a table 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 combinatory 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 good resolution and 600 for very good).
- "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 observer 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.

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

- "2D heat map apply ceil": apply a ceil to all heatmaps. This is good to compare all heatmaps with the same maximum. But, is very bad if some of them are flat 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 heatmap. There is a interface in Python that enables the user to rotate images. But, we didn't utilized it preferring a frame to display many heatmaps automatically, thus this 3D plot is static. It is divided in 5

color grades, from low (first mnat band, near zero) to high (maximum mutual information last band)
- "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:
  - (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.

16. Vertical Shannon Entropy

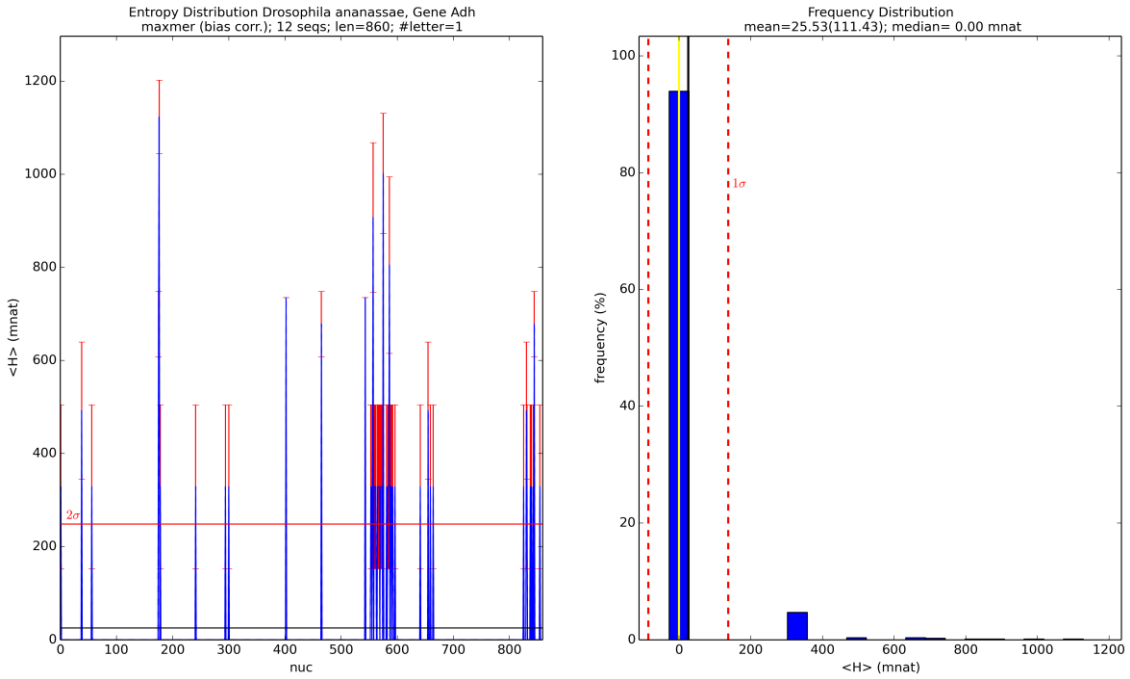

Figure 14 - VH distribution – Drosophila ananassae 12 sequences, L=860 nucleotides, NOL=1, Gene Adh, maxmer with bias correction

17. Vertical Mutual Information

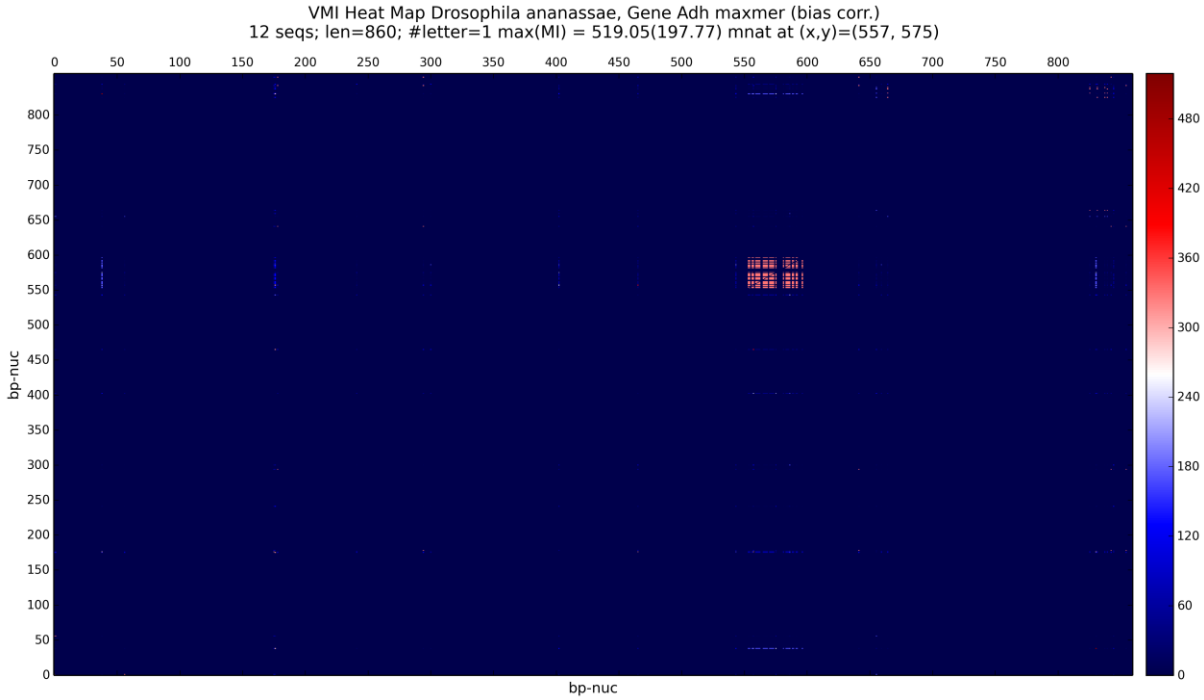

Figure 15 - VMI 2D heatmap – Drosophila ananassae 12 sequences, L=860 nucleotides, NOL=1, Gene Adh, maxmer with bias correction

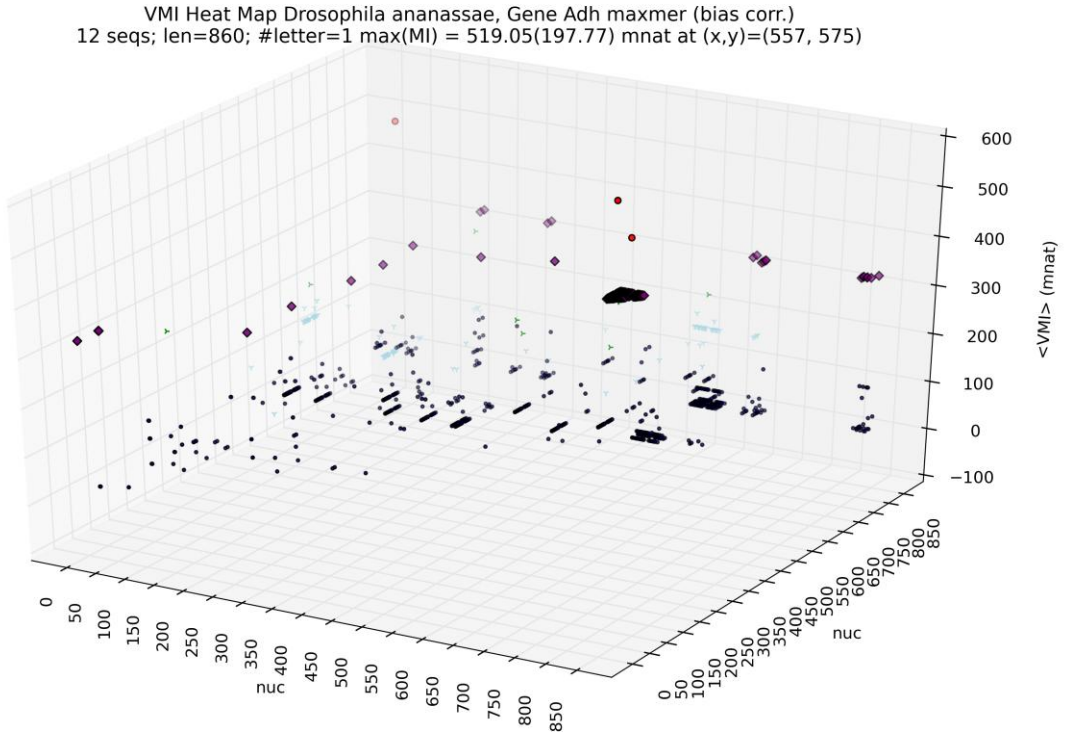

Figure 16 - VMI 3D heatmap – Drosophila ananassae 12 sequences, L=860 nucleotides, NOL=1, Gene Adh, maxmer with bias correction - this plot is generated from the same 2D heatmap data

18. 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 or **I** is a functional **VMI(i, j)** and is represented by a heatmap, given by,

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

where i and j are distinct positions and m and n are possible nucleotides ϵ {A,G,T,C}.

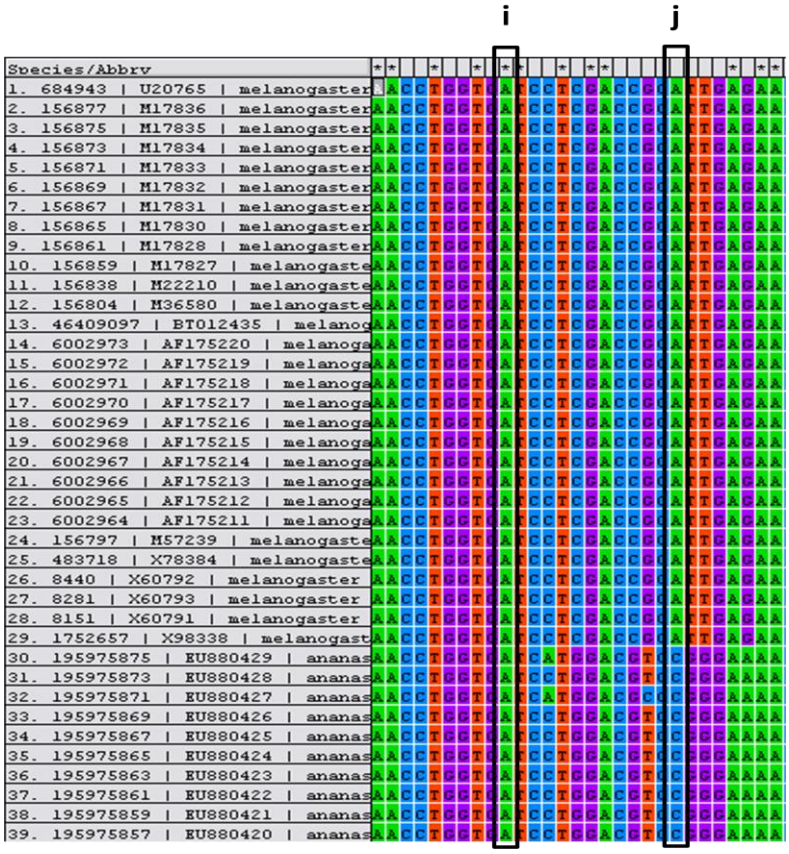

Figure 17 - Calculus of VMI cross-correlation – obtaining vertical $p_{mn}(i,j)$, $p_m(i,j)$ and $p_n(i,j)$, for j>i, and i,j = [1, L]

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

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

19. 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
- Mincut with bias correction
- Maxmer without bias correction
- Maxmer with bias correction

we decide to help the user with the button <calc all>. 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 this calculations may spent a lot of time (from a few minutes to a few hours 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.

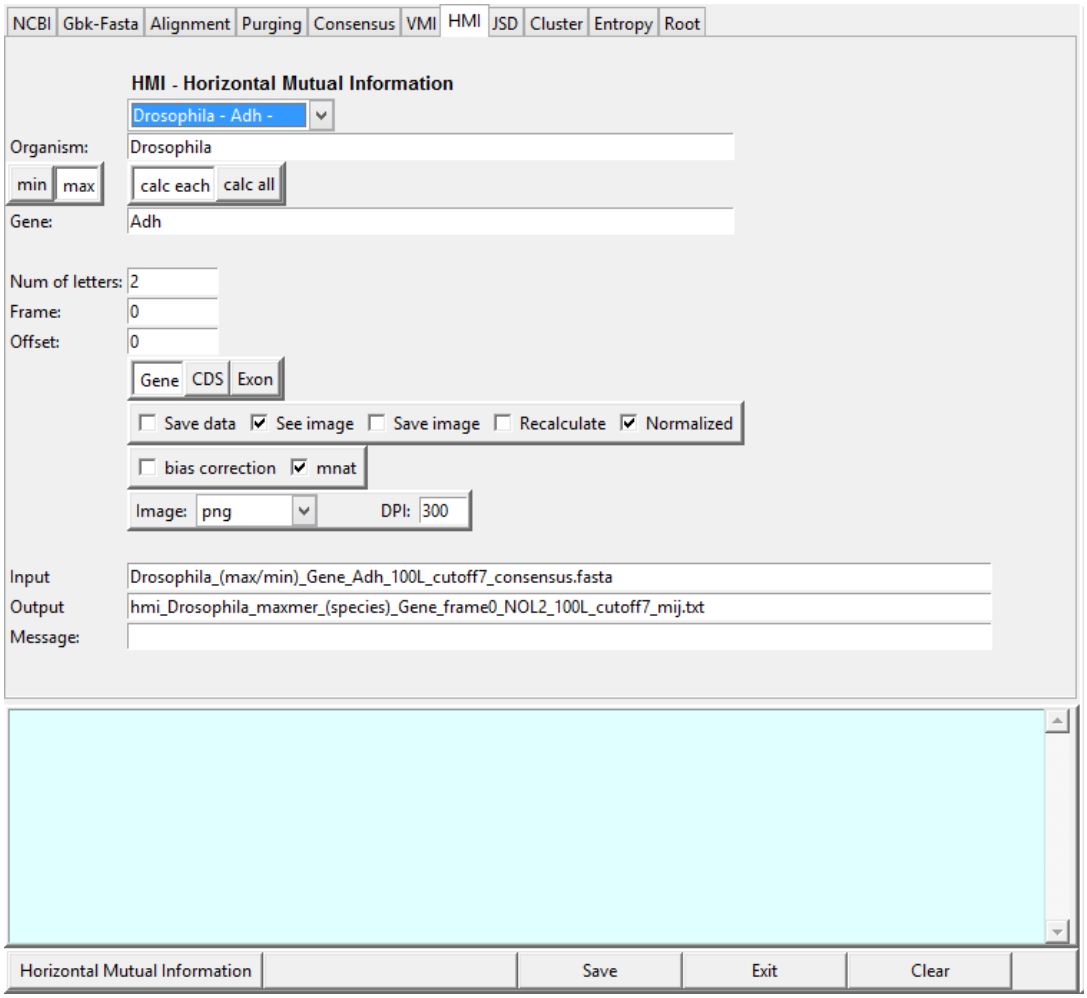

Figure 18 – Fifth Tab, Consensus, replace gaps by vertical consensus nucleotide

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 corretion").
- "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 in the end of all calculations. Summary file is a summary of all data as a table 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 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 good resolution and 600 for very good).
- "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 observer 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.

20. Horizontal Mutual Information

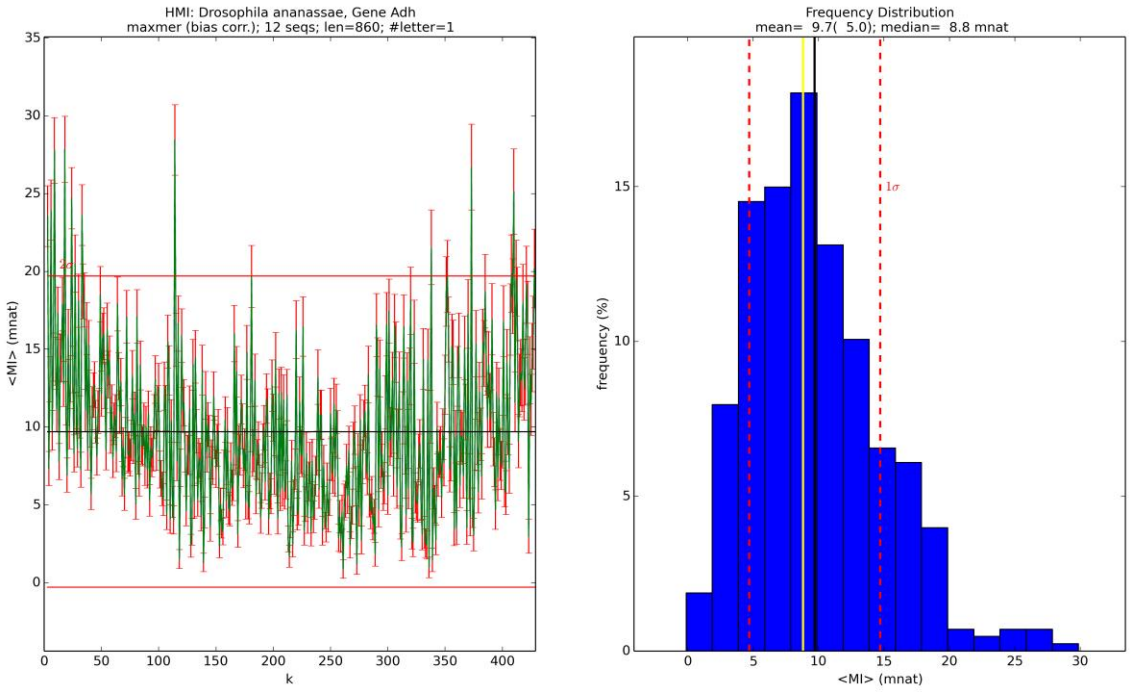

Figure 19 - HMI 2D heatmap – Drosophila ananassae 12 sequences, L=860 nuc, NOL=1, Gene Adh, maxmer with bias correction

21. 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 ϵ [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} ~< 10. In the present work we calculated **I (or HMI)** only for words with length equal 1 (one letter).

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

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

marginal probabilities are explained in VMI.

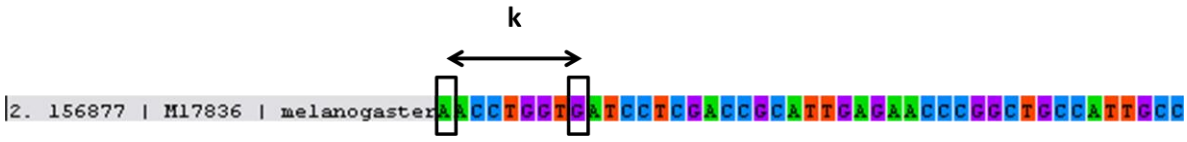

Figure 20 - How to calculate HMI autocorrelation – obtaining $p_{mn}(k)$ for k distance where m,n ϵ [A,G,T,C]

22. Jensen-Shannon Distance

Algorithm 08 (A8), tab JSD, is the algorithm that calculates the Jensen-Shannon Divergence (or Distance). JSD if a distance functional derived from Kullback-Leibler Divergence. It measures the distance between two distribution and is applied in MIA to discriminate species, in other words to measure the distance between species distribution pairwises.

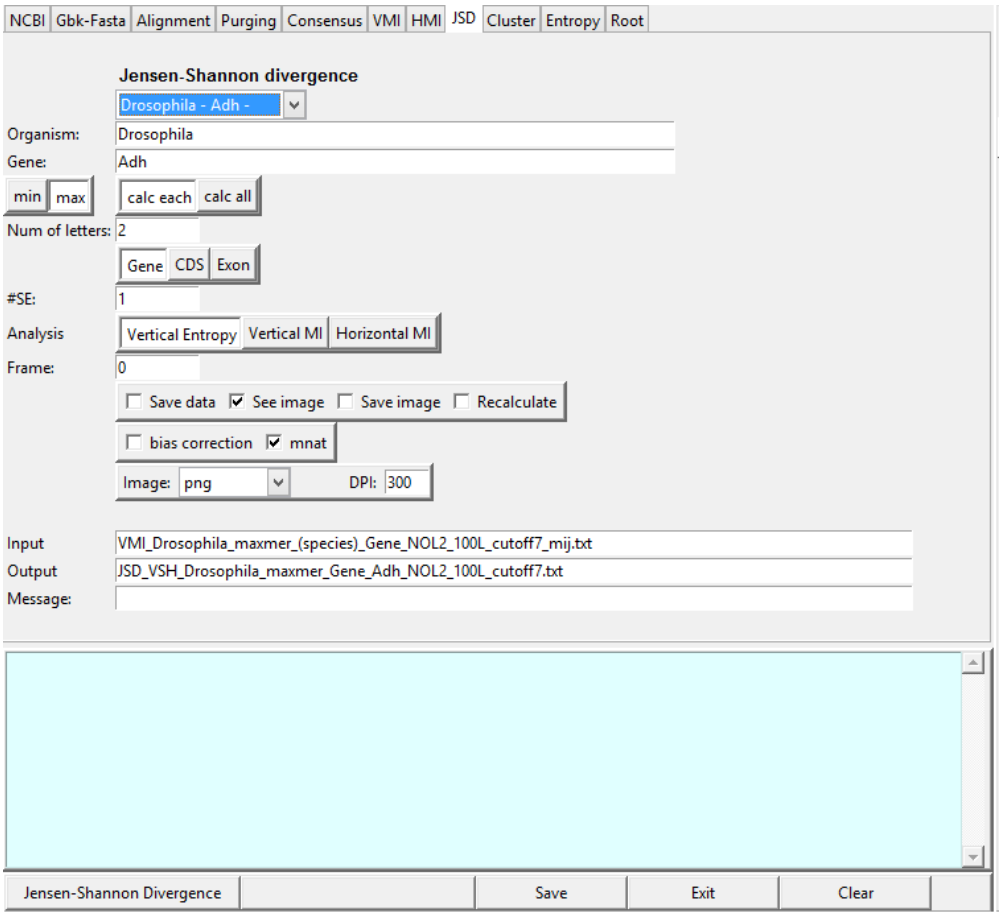

Figure 21 - JSD algorithm result in a table and a histogram with all distance between species distributions

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 corretion").
- "num of letter": you can vary NOL from 1 to any other integer (recommended < 10).
- "Vertical Entropy/VMI/HMI": choose which study 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 of all data as a table 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 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 good resolution and 600 for very good).
- "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.

23. JSD from Vertical Entropy

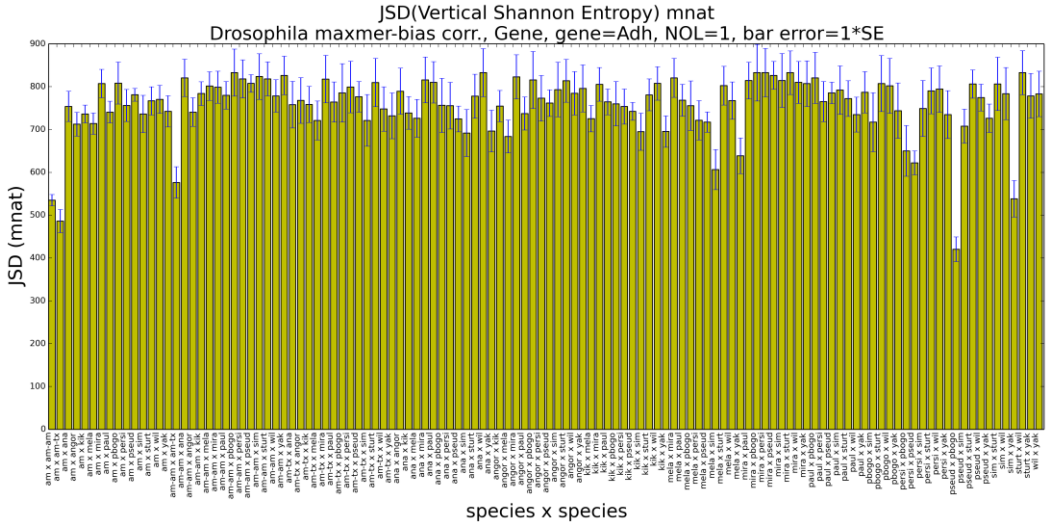

Figure 22a - JSD from Vertical Entropy, Gene Adh, maxmer with bias correction, NOL=1, SE = 1

24. JSD from Vertical Mutual Information

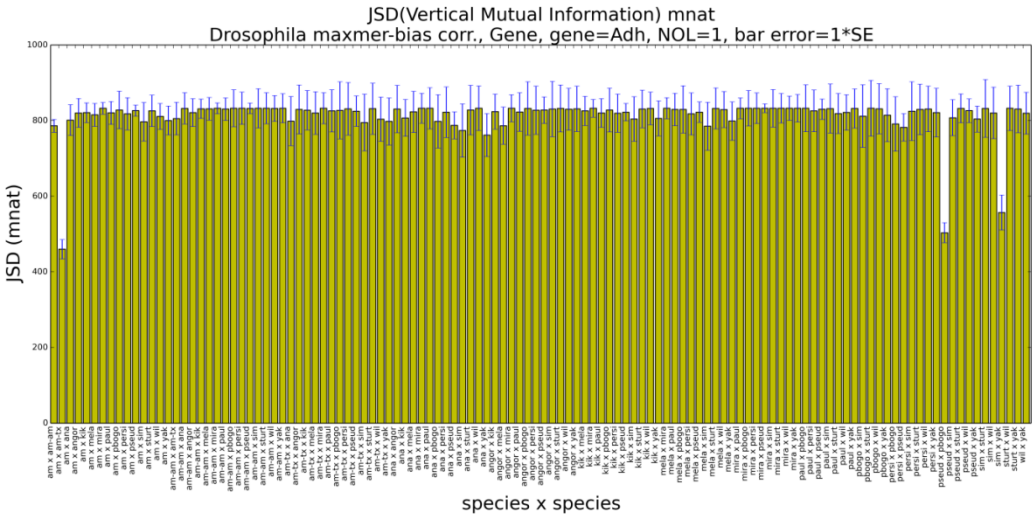

Figure 23b - JSD from Vertical Mutual Information, Gene Adh, maxmer with bias correction, NOL=1

25. JSD from Horizontal Mutual Information

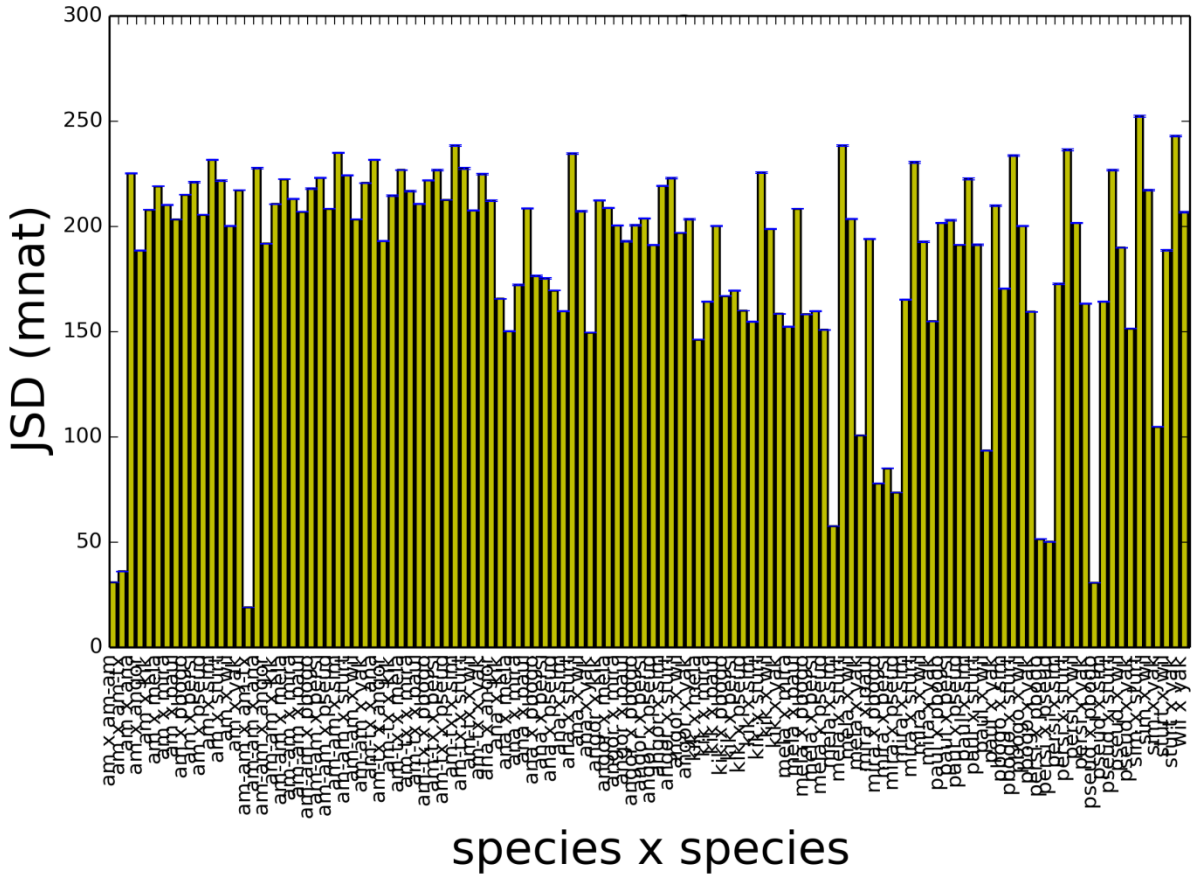

Figure 24 - JSD from Horizontal Mutual Information, Gene Adh, maxmer with bias correction, NOL=1

26. 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 normalized HMI or VMI, called HMI and VMI distributions. Thus, JSD is given by

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

27. 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 I_p+SE and I_p–SE x I_q+SE and I_q–SE, where p and q are different species indexes. Then we applied JSD to all four possible distributions and evaluate max(JSD) and min(JSD). The empirical SE is defined as:

$$SE[JSD(P||Q)] = (Max(JSD)-Min(JSD))/2 \quad (11)$$

28. 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.

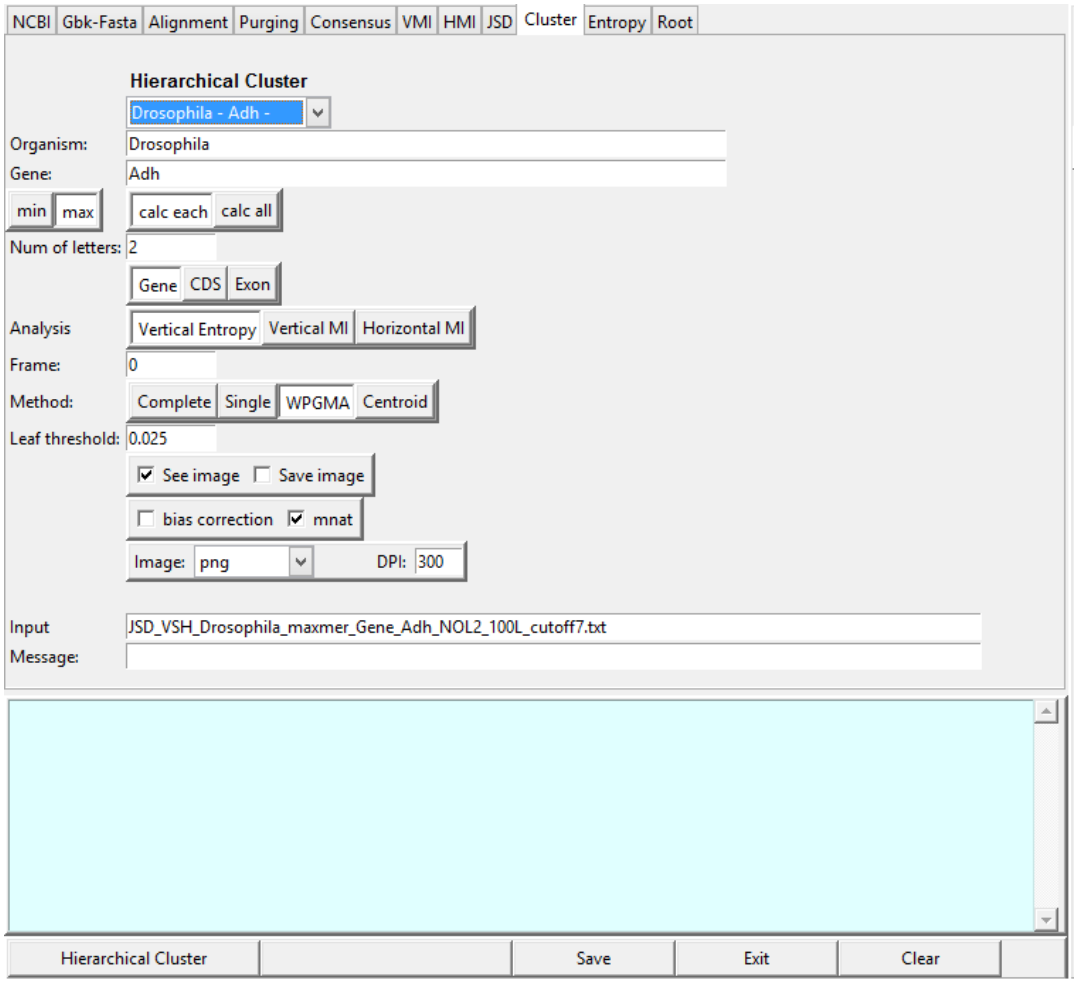

Figure 25 - Hierarchical cluster tab

Options:

- is almost the same as the lasts 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

29. Hierarchical Cluster Dendrogram

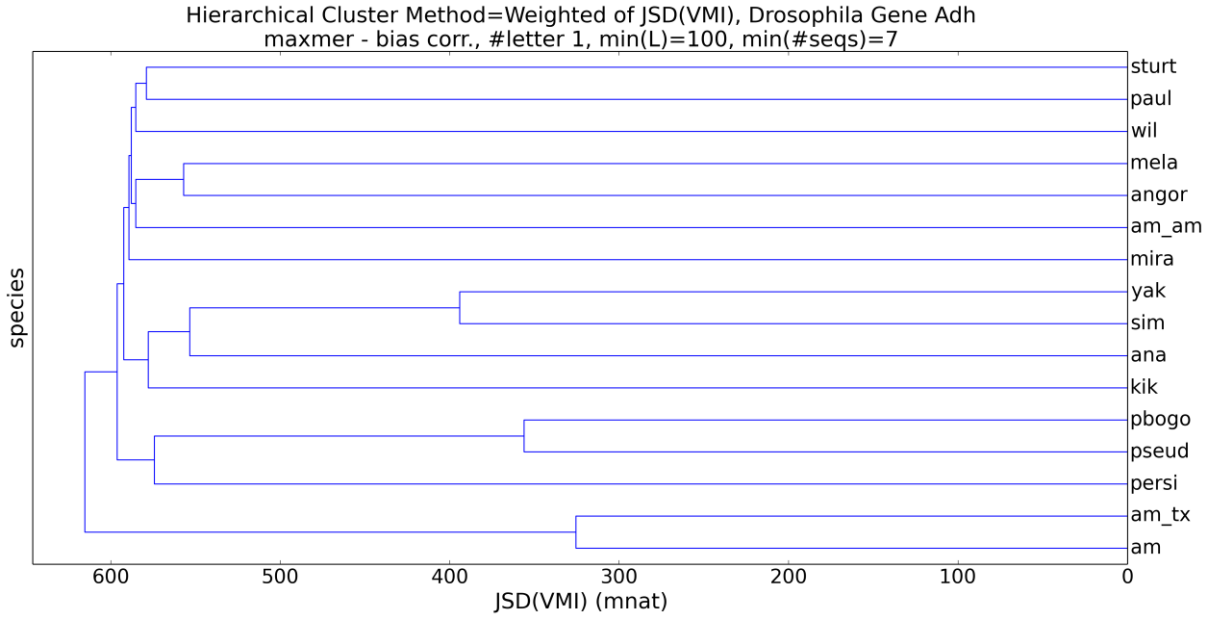

Figure 26 - Hierarchical cluster dendrogram from VMI, method=WPGMA, Gene Adh, maxmer with bias correction, NOL=1

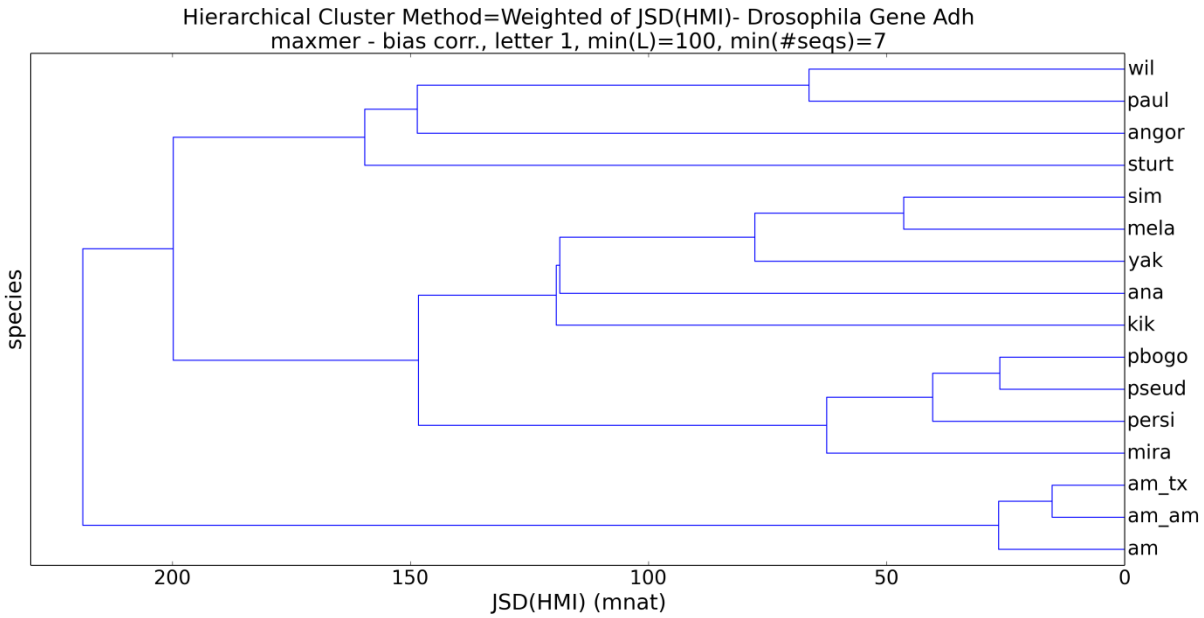

Figure 27 - Hierarchical cluster dendrogram from HMI, method=WPGMA, Gene Adh, maxmer with bias correction, NOL=1

30. Shannon Entropy (simulation)

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

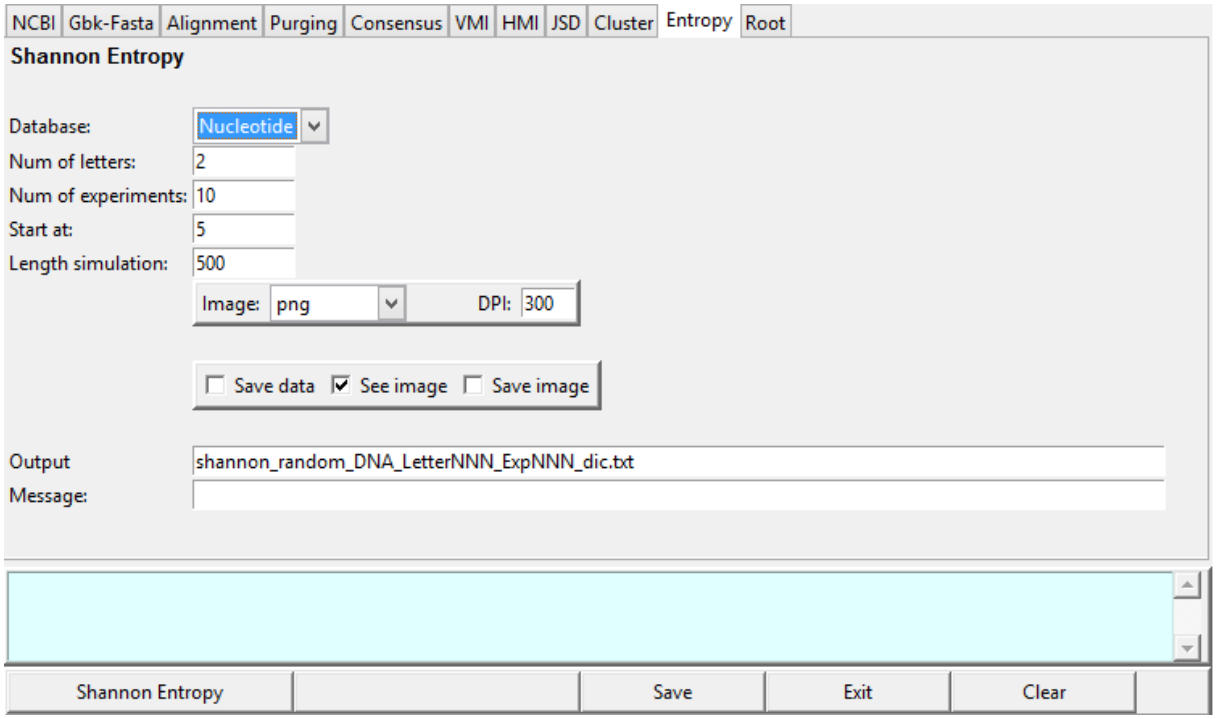

Figure 28 - Shannon Entropy Simulation

Options:

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

Run HC:

- click on <Hierarchical Cluster> to display the dendrogram.

- Input:
  - None
- Output:
  - An ASCII table in python dictionary format with name...
  - 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 to get a mean value an decrease the noise.

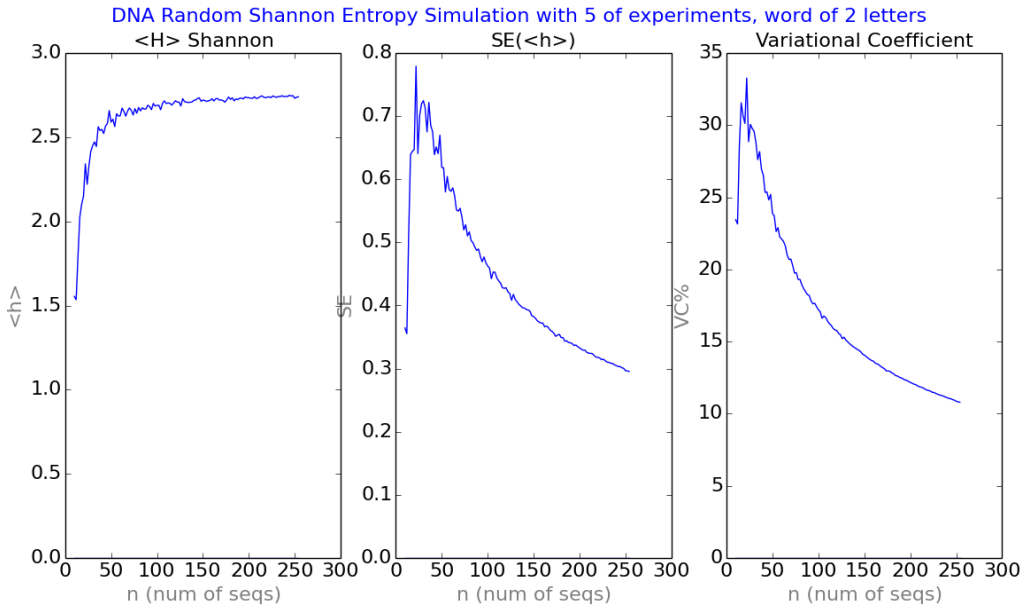

Figure 29 - Shannon Entropy Simulation, word of 2 letters, 5 repeated experiment results in high noise

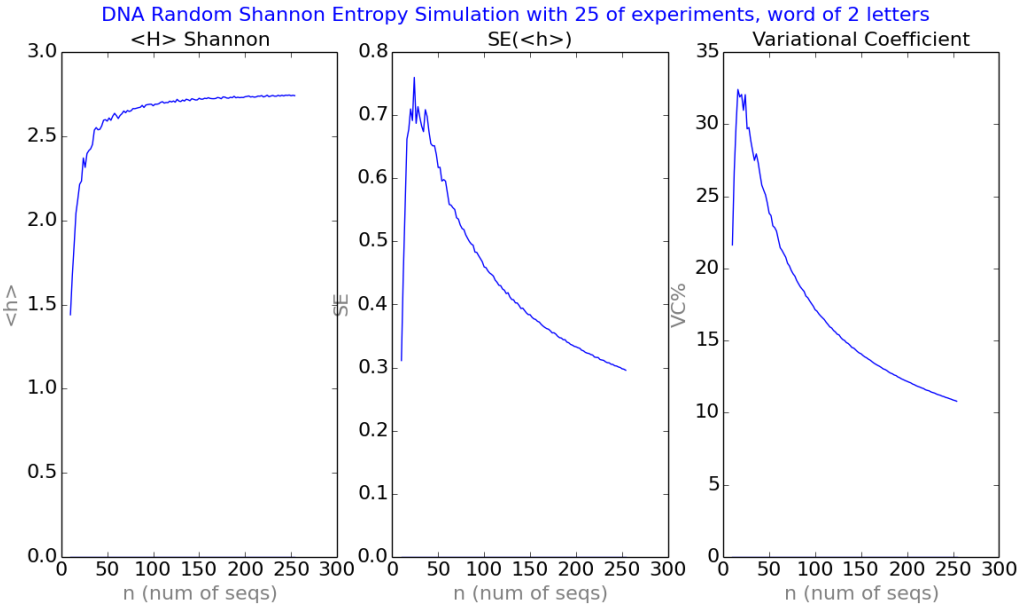

Figure 30 - Shannon Entropy Simulation, word of 2 letters, 25 repeated experiment results in medium noise

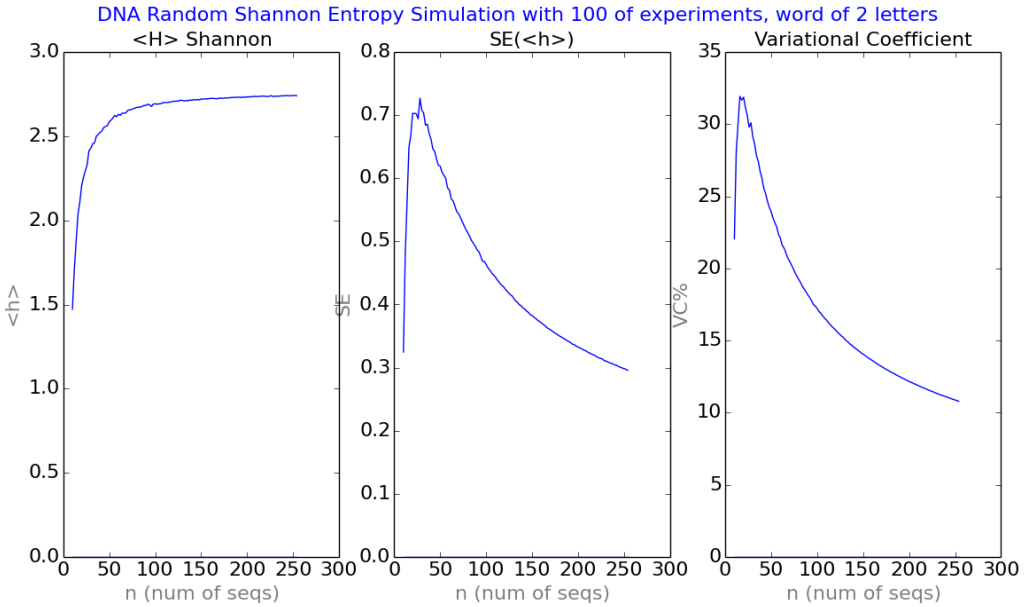

Figure 31 - Shannon Entropy Simulation, word of 2 letters, 100 repeated experiment results in low noise

31. 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.

Python Community (2012). Python Programming Language.

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