Reducing noise in protein multialignments

Marina Herrera Sarrias

Department of Mathematics, Stockholm University

mherrera@kth.se

January 6, 2019

1 Abstract

In this research a noise-reduction method is implemented in protein multialigments to evaluate its impact on phylogenic inference. The data used in this project is a reduced data set from the original data used by the creators of TrimAI. To test the data the program fastprot was used to obtain the distance matrices and fnj to infer a phylogenic tree for each alignment. The Python package DendroPy was used to measure the symmetric distance between the inferred tree (of the original and noise reduced alignment) and the reference tree. The results obtained showed that the symmetric difference between the reference tree and the noise reduced tree is in most of the cases lower than in the case of the symmetric distance between the reference tree and the original inferred tree. Although, in most of the cases the method applied did not make any change between the noise reduced inferred tree and the original inferred tree. This study focuses in determining whether the noise reduction makes any effect in the symmetric distance calculation.

2 Introduction

Phylogenetis is the study of the evolutionary relationship among a group of organisms. There is a strong relationship between the studies of phylogenetics and sequences analysis, if two sequences found in two different organisms are very similar we can then assume that they have derived from a common ancestor. When a multiple sequence alignment is performed it compares sequences in order to find similarities, it will firstly align the most similar pairs followed by the most distant pairs. The sequences alignment explains which positions are conserved from the ancestor sequence. Gaps in the alignments represent mutations therefore, the greater number of differences the more possible kind of mutations there will be.

Phylogenetic trees illustrate this relationship between organisms. The leaves of the tree represent the sequences, the nodes the relationship between them and the length of the branches, the number of changes that occurred in the sequences before the next level of separation. Trees could differ depending on the method used to construct them and there is not such as "best tree" whether there is a "good tree". The topology refers to the branching pattern of the tree, symmetry and asymmetry of a tree makes reference to how balanced or imbalanced the tree is, if the sequences(taxa) are evenly distributed among the clades the tree is balanced, otherwise is imbalanced. The objective of this project is to determine, whether it is worthwhile to use multialignment noise reduction in order to infeer a tree with fewer changes, and therefore a lower symmetric distance.

The reduced data set used in this project is composed by six different sub-directories, each sub-directory contains a reference tree and 300 alignments created by evolving sequences along the reference tree. Each pair of sub-directories present an average amount of mutations of 0.5, 1.0 and 2.0 per sequence site. Also, per each mutation rate there is one symmetric reference tree and one asymmetric reference tree.

In this project, the implemented noise-reduction method evaluates a multialigment column as noisy if there are more than 50% indels, at least 50% of the amino acids are unique and non of the amino acids appear more than twice. Each column takes the form of a sequence of amino acids given a position of the aligned sequences.

The results of the study showed that in most of the cases the applied noise reduction method did not have any effect on the symmetric distance calculation. Although it did showed a better performance in comparison with the original inferred tree, specially for those trees with higher average mutation rate.

3 Results and Discussion

After verifying the correct functionality of the program and processing the six data sets a frequency table of the six possible symmetric distance scenarios was compiled (Table 1).

In most of the cases the original tree and the noise reduced alignment tree are the same, indicating that the method applied had no effect in most of the alignments. In case for which the effect was not the same between the original and noise reduced inferred tree, the effect of reducing noise in the alignments makes the new noise reduced data to have a smaller symmetric distance with the reference tree, while there are fewer cases in which the effect is the opposite.

The reference tree was recovered with a greater frequency when the tree was symmetric and the average amount of mutations per sequence site was lower. From this, it can be understand that the symmetric difference is equal to 0. According to the frequency values, the reference tree was recovered 3 times by the noise reduced tree while it was recovered 5 times by the original multi-alignment inferred tree.

Table 2 gives statistical information about the six data sets. Both inferred trees (original and noise reduced) take nearly the same minimum (not taking into account distances of 0), maximum, mean and median values.

Table 1. Frequencies of the symmetric distances of the original and reduced newick trees compared to the reference tree.

	Asymm_0.5	Asymm_1.0	Asymm_ 2.0	Symm_0.5	Symm_1.0	Symm_2.0
Original >Reduced	45	60	79	44	59	83
Original <Reduced	48	54	45	28	36	54
Original = Reduced	206	186	176	214	201	162
Original = 0	0	0	0	4	1	0
Reduced = 0	0	0	0	2	1	0
Original & Red = 0	1	0	0	8	2	1
Total	300	300	300	300	300	300

Table 2. Data sets statistics referring to the symmetric distance of the original and reduced newick trees with the reference tree.

	Minimum		Maximum		Mean		Median	
	Original	Reduced	Original	Reduced	Original	Reduced	Original	Reduced
Asymmetric 0.5	2	2	18	18	7.63	7.62	8	8
Asymmetric 1.0	2	2	18	18	9.43	9.39	10	10
Asymmetric 2.0	4	4	22	20	12.34	11.99	12	12
Symmetric 0.5	2	2	10	10	4.66	4.55	4	4
Symmetric 1.0	2	2	12	12	6.23	6.05	6	6
Symmetric 2.0	2	2	16	16	8.91	8.65	8	8

The noise reduction performance was slightly higher on the symmetric sets compared to the asymmetric (Table 3) , the noise reduction average was observed to increase with the average mutation rate.

Table 3. Average noise reduction performed per data set

	Average Noise Reduction
Asymmetric 0.5	0.146
Asymmetric 1.0	0.199
Asymmetric 2.0	0.252
Symmetric 0.5	0.172
Symmetric 1.0	0.221
Symmetric 2.0	0.282

4 Methods and Materials

Three different major tasks were first developed separately, and later joined together to consolidate the program and conduct an exploratory data analysis. The individual scripts were initially tested on the experimental data and control cases and then merged in a single script to later run the program with the fixed data set.

A Github repository was created in order to control the program's version. The Laboratory notebook was created using a Jupyter notebook Markdown and later compiled into a HTML file, which can be found on the Github repository. This notebook contains all details regarding the development of the current project.

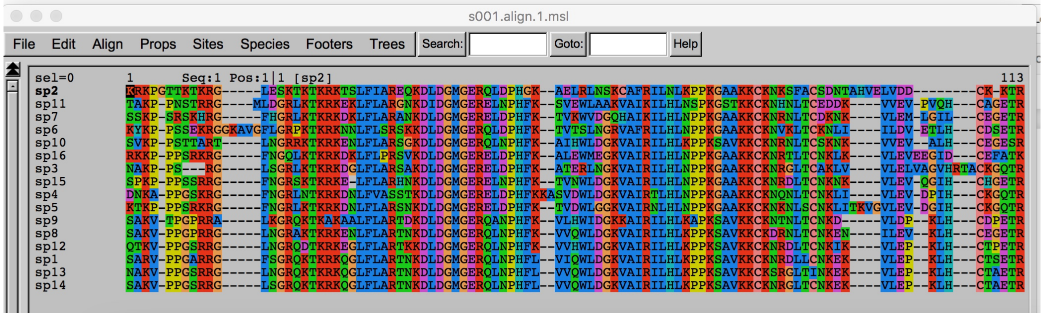
Following the course centerpiece guideline, controls were introduced along the project as well error handling. When an error occurs the program aborts and prints a message to standard error. All standard-in files are read by robust libraries. Moreover, a sub-directory containing a subset of the fixed data set was created in order to test the program, controls for different scenarios were also introduced to ensure the program functionality. No input is required for testing the data, neither for running the main program.

4.1 Running: run_program.py

The python script run_program.py performs all the required tasks of the different stages of this project:

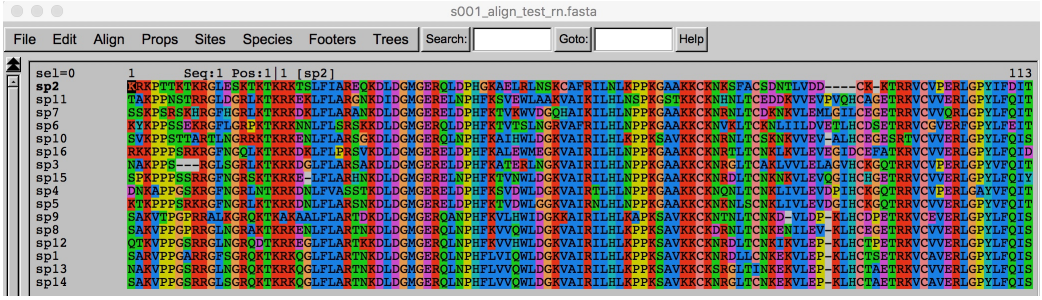
1. Reduce protein multialignment noise: The program first takes an alignment from standard-in, filters the columns and writes a new file containing the noise reduced multi-alignment. In order to illustrate the functionality of the program, Seaview was used as a graphical interface for the following example.

   As shown below, the original alignment before (Figure 1) and after performing noise reduction (Figure 2). In this particular case the reduced noise represented a 17.08% of the original sequence.



```
sel=0     1      Seq:1 Pos:1|1 [sp2]                                                              113
sp2       KRKPGTTKTKRG-----LESKTKTKRKTSLFIAREQKDLDGMGERQLDPHGK--AELRLNSKCAFRILNLKPPKGAAKKCKNKSFACSDNTAHVELVDD--------CK-KTR
sp11      TAKP-PNSTRRG----MLDGRLKTKRKEKLFLARGNKDIDGMGERELNPHFK--SVEWLAAKVAIKILHLNSPKGSTKKCKNHNLTCEDDK----VVEV-PVQH---CAGETR
sp7       SSKP-SRSKHRG-----FHGRLKTKRKDKLFLARANKDLDGMGERELDPHFK--TVKWVDGQHAIKILHLNPPKGAAKKCKNRNLTCDKNK----VLEM-LGIL---CEGETR
sp6       KYKP-PSSEKRGGKAVGFLGRPKTKRKNNLFLSRSKKDLDGMGERQLDPHFK--TVTSLNGRVAFRILHLNPPKGAAKKCKNVKLTCKNLI----ILDV-ETLH---CDSETR
sp10      SVKP-PSTTART-----LNGRRKTKRKENLFLARSGKDLDGMGERQLNPHFK--AIHWLDGKVAIRILHLKPPKSAVKKCKNRNLTCSKNK----VVEV--ALH---CEGESR
sp16      RKKP-PPSRKRG-----FNGQLKTKRKDKLFLPRSVKDLDGMGERELDPHFK--ALEWMEGKVAIRILHLNPPKGAAKKCKNRTLTCNKLK----VLEVEEGID---CEFATK
sp3       NAKP-PS---RG-----LSGRLKTKRKDGLFLARSAKDLDGMGERELDPHFK--ATERLNGKVAIRILHLNPPKGAAKKCKNRGLTCAKLV----VLELVAGVHRTACKGQTR
sp15      SPKP-PPSSRRG-----FNGRSKTKRKE-LFLARHNKDLDGMGERELNPHFK--TVNWLDGKVAIRILHLNPPKGAAKKCKNRDLTCNKNK----VLEV-QGIH---CHGETR
sp4       DNKA-PPGSKRG-----FNGRLNTKRKDNLFVASSTKDLDGMGERELDPHFKKASVDWLDGKVAIRTLHLNPPKGAAKKCKNQNLTCNKLI----VLEV-DPIH---CKGQTR
sp5       KTKP-PPSRKRG-----FNGRLKTKRKDNLFLARSNKDLDGMGERELDPHFK--TVDWLGGKVAIRNLHLNPPKGAAKKCKNKNLSCNKLITKVGVLEV-DGIH---CKGQTR
sp9       SAKV-TPGPRRA-----LKGRQKTKAKAALFLARTDKDLDGMGERQANPHFK--VLHWIDGKKAIRILHLKAPKSAVKKCKNTNLTCNKD-----VLDP--KLH---CDPETR
sp8       SAKV-PPGPRRG-----LNGRAKTKRKENLFLARTNKDLDGMGERQLNPHFK--VVQWLDGKVAIRILHLKPPKSAVKKCKDRNLTCNKEN----ILEV--KLH---CEGETR
sp12      QTKV-PPGSRRG-----LNGRQDTKRKEGLFLARTKKDLDGMGERQLNPHFK--VVHWLDGKVAIRILHLKPPKSAVKKCKNRDLTCNKIK----VLEP--KLH---CTPETR
sp1       SARV-PPGARRG-----FSGRQKTKRKQGLFLARTNKDLDGMGERQLNPHFL--VIQWLDGKVAIRILHLKPPKSAVKKCKNRDLLCNKEK----VLEP--KLH---CTSETR
sp13      NAKV-PPGSRRG-----LNGRQKTKRKQGLFLARTNKDLDGMGERQLNPHFL--VVQWLDGKVAIRILHLKPPKSAVKKCKSRGLTINKEK----VLEP--KLH---CTAETR
sp14      SAKV-PPGSRRG-----LSGRQKTKRKQGLFLARTNKDLDGMGERQLNPHFL--VVQWLDGKVAIRILHLKPPKSAVKKCKNRGLTCNKEK----VLEP--KLH---CTAETR
```

Figure 1: Original s001.align.1.msl



```
sel=0     1      Seq:1 Pos:1|1 [sp2]                                                              113
sp2       KRKPTTKTKRGLESKTKTKRKTSLFIAREQKDLDGMGERQLDPHGKAELRLNSKCAFRILNLKPPKGAAKKCKNKSFACSDNTLVDD----CK-KTRRVCVPERLGPYIFDIT
sp11      TAKPPNSTRRGLDGRLKTKRKEKLFLARGNKDIDGMGERELNPHFKSVEWLAAKVAIKILHLNSPKGSTKKCKNHNLTCEDDKVVEVPVQHCAGETRKVCVVERLGPYLFQIT
sp7       SSKPSRSKHRGFHGRLKTKRKDKLFLARANKDLDGMGERELDPHFKTVKWVDGQHAIKILHLNPPKGAAKKCKNRNLTCDKNKVLEMLGILCEGETRRVCVVQRLGPYLFQIT
sp6       KYKPPSSEKRGFLGRPKTKRKNNLFLSRSKKDLDGMGERQLDPHFKTVTSLNGRVAFRILHLNPPKGAAKKCKNVKLTCKNLIILDVETLHCDSETRRVCGVERFGPYLFEIT
sp10      SVKPPSTTARTLNGRRKTKRKENLFLARSGKDLDGMGERQLNPHFKAIHWLDGKVAIRILHLKPPKSAVKKCKNRNLTCSKNKVVEV-ALHCEGESRTVCVVERLGPYLFQIS
sp16      RKKPPPSRKRGFNGQLKTKRKDKLFLPRSVKDLDGMGERELDPHFKALEWMEGKVAIRILHLNPPKGAAKKCKNRTLTCNKLKVLEVEGIDCEFATKRVCVVERLGPYLFQID
sp3       NAKPPS---RGLSGRLKTKRKDGLFLARSAKDLDGMGERELDPHFKATERLNGKVAIRILHLNPPKGAAKKCKNRGLTCAKLVVLELAGVHCKGQTRRVCVPERLGPYVFQIT
sp15      SPKPPPSSRRGFNGRSKTKRKE-LFLARHNKDLDGMGERELNPHFKTVNWLDGKVAIRILHLNPPKGAAKKCKNRDLTCNKNKVLEVQGIHCHGETRRVCVVERLGPYLFQIY
sp4       DNKAPPGSKRGFNGRLNTKRKDNLFVASSTKDLDGMGERELDPHFKSVDWLDGKVAIRTLHLNPPKGAAKKCKNQNLTCNKLIVLEVDPIHCKGQTRRVCVPERLGAYVFQIT
sp5       KTKPPPSRKRGFNGRLKTKRKDNLFLARSNKDLDGMGERELDPHFKTVDWLGGKVAIRNLHLNPPKGAAKKCKNKNLSCNKLIVLEVDGIHCKGQTRRVCVVERLGPYVFQIT
sp9       SAKVTPGPRRALKGRQKTKAKAALFLARTDKDLDGMGERQANPHFKVLHWIDGKKAIRILHLKAPKSAVKKCKNTNLTCNKD-VLDP-KLHCDPETRKVCEVERLGPYLFQIS
sp8       SAKVPPGPRRGLNGRAKTKRKENLFLARTNKDLDGMGERQLNPHFKVVQWLDGKVAIRILHLKPPKSAVKKCKDRNLTCNKENILEV-KLHCEGETRRVCVVERLGPYLFQIS
sp12      QTKVPPGSRRGLNGRQDTKRKEGLFLARTKKDLDGMGERQLNPHFKVVHWLDGKVAIRILHLKPPKSAVKKCKNRDLTCNKIKVLEP-KLHCTPETRKVCVVERLGPYLFQIS
sp1       SARVPPGARRGFSGRQKTKRKQGLFLARTNKDLDGMGERQLNPHFLVIQWLDGKVAIRILHLKPPKSAVKKCKNRDLLCNKEKVLEP-KLHCTSETRKVCAVERLGPYLFQIS
sp13      NAKVPPGSRRGLNGRQKTKRKQGLFLARTNKDLDGMGERQLNPHFLVVQWLDGKVAIRILHLKPPKSAVKKCKSRGLTINKEKVLEP-KLHCTAETRKVCVVERLGPYLFQIS
sp14      SAKVPPGSRRGLSGRQKTKRKQGLFLARTNKDLDGMGERQLNPHFLVVQWLDGKVAIRILHLKPPKSAVKKCKNRGLTCNKEKVLEP-KLHCTAETRKVCAVERLGPYLFQIS
```

Figure 2: s001.align.1.msl after noise reduction

2. Generate phylogenetic trees: On the second stage, the program pars through all the alignments available in the data directory (original and noise reduced), and using the fastphylo programs (fastprot and fnj) computes a distance matrix and generate a phylogenetic tree per alignment. In order to accomplish this task the program uses a sub-process to pipe the matrix creation and generate the tree. As the symmetric distance (computed in next stage) only requires the tree's topology it will not be necessary to compute the branch length information.

   Following the s001.align.1.msl alignment example both, the noise reduced and original alignment have the same phylogenetic tree shown in Figure 3 while the reference tree is represented by Figure 4.

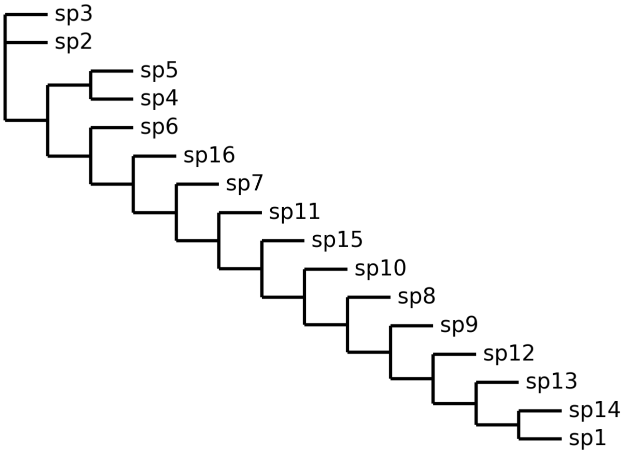


Figure 3: s001.align.1.msl original and noise reduced tree

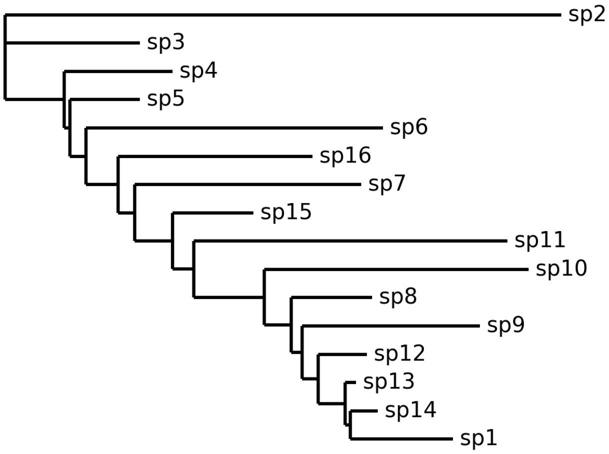


Figure 4: asymmetric_0.5.tree reference tree

3. Compute trees distances: Finally, the program computes the symmetric distance between the reference tree and both, original and noise reduced tree using the python's module Dendropy. The symmetric distance is a count of how many partitions are between the two trees, that are present in one tree but not in the other. Following the s001.align.1.msl, alignment example, as both the original tree and noise reduced tree are the same the distance with the reference tree will be therefore the same, in this case 4.

4.2 run_program.py: Error handling and Control cases

The script run_program.py will abort if any of the following cases occur:

- There is an input introduction when executing the program.
- The data directory is empty.
- The data sub-directories are empty. (e.g asymmetric_0.5).
- Any of the standard-in file (.msl,.tree) is empty.
- All the columns in the alignment are noisy.

On the other hand, the program will print a warning message but will not abort processing if:

- The alignment do not have noisy columns.
- The noise reduction rate is greater than 0.5 times the original alignment.

All of the control files used for testing the accuracy of this script can be found in the results directory of Github repository, in which a file for each case above-mentioned can be found.