Reducing Noise in Protein Multiple Sequence Alignments

Annika Bendes, Hanna Hassan
DD2404 Applied Bioinformatics

Abstract

Multiple sequence alignments, also known as MSAs, are alignments of three or more biological sequences. MSAs can be used to infer sequence homology and these sequences' shared evolutionary origins can be assessed through phylogenetic analysis. Here we have developed a noise reduction program for MSAs based on the criteria that a column in the alignments is considered noisy if there are more than 50% indels, if at least 50% of amino acids are unique, or if no amino acids appear more than twice. Furthermore, we have evaluated the program's performance by using symmetric differences between the phylogenetic tree for the MSAs and a reference tree, and investigated whether it is worthwhile to use multialignment noise reduction based on the criteria above. The evaluation of the program was performed using a reduced dataset of the data used by the authors of TrimAl in their evaluation. We obtained an insignificant improvement of the symmetric difference for the noise reduced alignments compared to the unmodified alignments with regard to the reference, indicating that it is not worthwhile to use multialignment noise reduction based on the given criteria.

Introduction

Multiple sequence alignments (MSA) are alignments of three or more biological sequences. In the field of Bioinformatics, MSAs are used to infer homology between sequences. From the MSA, the shared evolutionary origins of the sequences can be assessed through phylogenetic analysis. However, biological sequences can contain regions which are not inherited which can cause problems if aligned in an MSA since this would affect the phylogenetic analysis. Fast accumulation of mutations can also infer problems when determining an evolutionary relationship if the sequences cannot be correctly aligned due to large differences between the mutated sequence and its common ancestors. In order to reduce these problems it could be helpful to remove such problematic regions, referred to as noisy, from the aligned sequences.

In this project, a simple noise-reduction method was implemented and its impact on the phylogeny inference was assessed. The program for reducing noise in MSAs was written in Python 2.7 and packages such as *BioPython*, *FastPhylo* and *DendroPy* were used.

Materials and Methods

Test data
This project used a reduced dataset of the data which was originally used by the authors of TrimAl in their evaluation. The dataset includes six subsets divided into two different categories: symmetric and

asymmetric. Each category has three subsets which contain one tree file, a reference tree, and 300 alignments created by evolving sequences along the reference tree and then aligning the results. The following list shows how the three subsets used in each category are categorized into the average number of mutations per site in the sequences.

I. Symmetric

- `symmetric_0.5 (0.5 mutations per site)`
- `symmetric_1.0 (1 mutations per site)`
- `symmetric_2.0 (2 mutations per site)`

II. Asymmetric

- `asymmetric_0.5 (0.5 mutations per site)`
- `asymmetric_1.0 (1 mutations per site)`
- `asymmetric_2.0 (2 mutations per site)`

Noise reduction

In this project, a column in a multi-alignment is considered noisy if any of the following criterias are fulfilled:

- there are more than 50% indels
- at least 50% of amino acids are unique
- no amino acid appears more than twice

The Python program written in this project takes a multi-alignment as input and as a first step removes the columns which fulfill at least one criteria shown above which is referred to as noise reduction. How the noise reduction affects the multi-alignment is evaluated through the steps mentioned below.

1. A tree is inferred from the given multi-alignment
2. The multi-alignment is noise reduced
3. A tree is inferred from the noise reduced multi-alignment
4. Both trees are compared to the reference tree by their respective symmetric difference.

If the symmetric difference to the reference tree is smaller after noise reduction, the reduction is considered successful as the tree is now closer to the reference tree than the original. If there is no difference, the trees are considered identical. If the difference is bigger between the noise reduced tree and the reference tree, the noise reduction did not perform well.

Project setup

The code building up the program is comprised of two scripts. One of the scripts, *reducenoise.py*, contains functions that reduce the noise in a given multi-alignment according to the criteria stated above. The other script, *huvudprog.py*, executes the other script's functions for the available test data. The script *huvudprog.py* takes a subset of the data consisting of a number of MSAs in the form of FASTA files and one reference tree, and for each FASTA input file it infers an unweighted tree using the functions *fastprot* and *fnj* from *fastphylo* package. Thereafter, the scripts creates a noise-reduced version of the MSA using

reducenoise.py and infers a second tree for the noise reduced MSA. The two trees are then compared with the reference tree using *DendroPy*. The output from the program is the symmetric differences between the trees and the reference trees, as well as the difference between the alignment trees and the noise reduced trees in forms of CSV files. The output is written to a */results* directory. The obtained CSV files were later analyzed using Microsoft Excel.

During the process of writing the program, almost all of Stafford Noble's ideas were implemented. E.g. the files were separated into folders in a top-level chronological directory structure, a lab notebook was kept and most rules of thumb for carrying out an experiment were used. However, version control software was not used due to unfamiliarity with such software. A backup of the script was instead created by using external hard drives and cloud-services. Files were not overwritten on, changes that were made were documented and the script was never worked on simultaneously.

Controls

To ensure that the written program performs well and detects errors, test cases were set up. These were set up to see how the program handles different types of data as input and to implement controls in order to obtain results that are as correct as possible. These tests involved cases such as using empty, small and null data as input. Input alignments with an improper format as well as where the output was known beforehand were also used to ensure correctness. To handle errors, messages are printed to standard error.

Results and Discussion

In order to evaluate the impact of our simple noise-reduction program on phylogeny inference we inferred phylogenetic trees for the original alignments and the noise reduced alignments, and compared these trees to reference trees. In average, the modified alignments have a lower symmetric difference to the reference trees compared to the unmodified alignments, showing that these generated trees are closer to the reference (Figure 1). However, the difference is very small and lies within the standard deviation for the data sets, and is not significant. The results indicates that noise reduction may be more effective the more the test data was mutated (Table 1), but the test data did not contain sequences that had more than 2 mutations per site. Analyzes using data sets with more mutations would be necessary to evaluate this further. The frequency of recovery of the reference tree was higher for the symmetric data compared to the asymmetric data for both the modified and unmodified alignments. But the difference between the symmetric difference for the modified and unmodified alignments were not significant (Table 2).

Conclusion

The insignificant improvement of the modified alignments compared to the unmodified alignments with regard to the reference, and the high failure rate of the modified alignments being closer to the reference tree than the unmodified alignments indicates that it is not worthwhile to use multialignment noise reduction based on the given criteria.

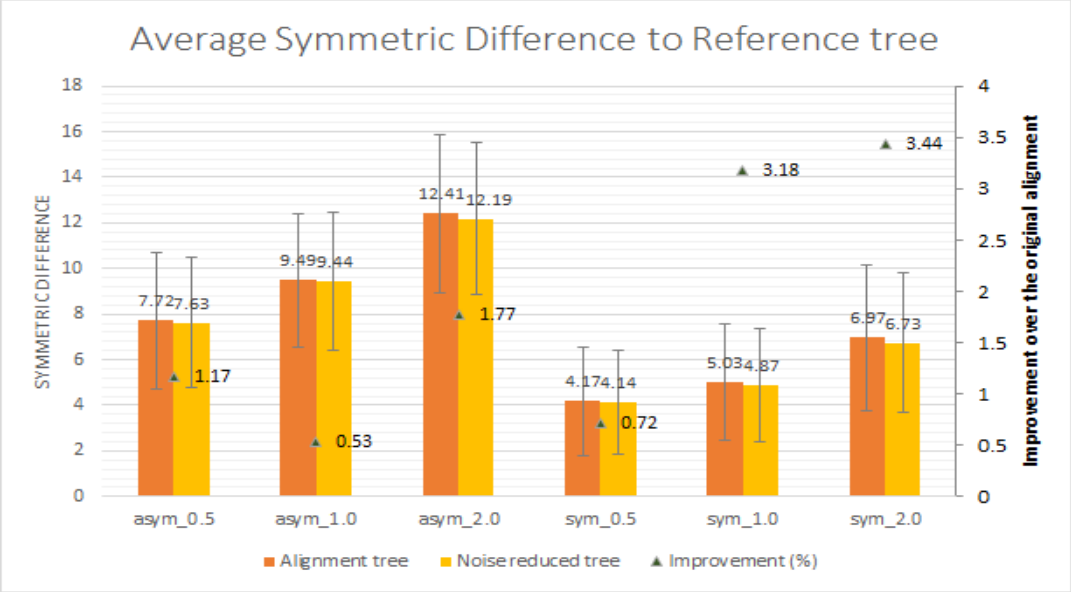

Figure 1 Average symmetric difference between trees and the reference tree. The standard deviation is shown as error bars.

Table 1 *Failure and Success Rate of noise reduction, where failure means that the symmetric difference between the noise reduced tree and the reference tree are greater than the symmetric difference between the unmodified tree and the reference tree, and vice versa.*

	asym_0.5	***asym_1.0***	***asym_2.0***	***sym_0.5***	***sym_1.0***	***sym_2.0***
Less accurate (Failure)	*25*	*42*	*46*	*30*	*36*	*48*
	(8.3%)	*(14.0%)*	*(15.3%)*	*(10.0%)*	*(12.0%)*	*(16.0%)*
More accurate (Success)	*33*	*50*	*72*	*35*	*53*	*73*
	(11.0%)	*(16.7%)*	*(24.0%)*	*(11.7%)*	*(17.7%)*	*(24.3%)*
Same results	*242*	*208*	*182*	*235*	*211*	*179*
	(80.7%)	*(69.3%)*	*(60.7%)*	*(78.3%)*	*(70.3%)*	*(59.7%)*

Table 2 *Frequency of reference tree being recovered*

	Before noise reduction	*After noise reduction*
asymmetric_0.5	*1*	*1*
asymmetric_1.0	*0*	*0*
asymmetric_2.0	*0*	*0*
symmetric_0.5	*23*	*18*
symmetric_1.0	*16*	*18*
symmetric_2.0	*4*	*5*