Evaluating Noise Reduction in Protein Multialignments

Filip Domazet

Introduction

Protein multialignments are alignments of three or more protein sequences. Such alignments are performed in order to infer sequence homology and conduct a phylogenetic analysis to assess the sequences' shared evolutionary origin.

These alignments can be noisy as some regions in the protein are not inherited and others could have evolved so quickly that the correct alignment is impossible to infer. In order to get better results in subsequent analyses one could try avoiding these regions by removing such "bad" columns from the multialignment. This project aims to evaluate whether performing such noise reduction causes significant improvement by comparing trees inferred from artificially generated data to their reference trees.

Results

In order to test whether noise reduction is worthwhile, six data sets were used. Each data set was comprised of 300 protein multiple sequence alignments from sequences which were artificially evolved from the reference tree. Three of the sets have a symmetric reference tree, and three have an asymmetric one. Both cases use three different mutation rates, ie. the average number of mutations per site. These are 0.5, 1.0 and 2.0.

For each such alignment, another alignment was generated but with removed "bad" columns. For both of these, a distance matrix between sequences is calculated and a tree is then inferred. The last step was to calculate the distance to the reference tree.

Once this has been performed for all the alignments in a data set, it was evaluated whether the tree inferred from the alignment with reduced noise shows a significantly smaller distance to the reference tree. It was also evaluated whether the distance is significantly larger, in order to also verify whether performing noise reduction actually causes worse results. The results are shown in Table 1, with a significance level of 5% being used.

Alternative hypothesis	Ref tree type:	Symmetric			Asymmetric		
	Mutation rate:	0.5	1.0	2.0	0.5	1.0	2.0
$\mu_{noisy} > \mu_{denoised}$	Significance:	yes	no	yes	no	no	no
$\mu_{noisy} < \mu_{denoised}$	Significance:	no	no	no	no	no	yes

Table 1 - Significance of distance to reference tree difference

Overall, the results do not seem to encourage performing noise reduction. Out of six tests, only two show a significant improvement when using alignments with reduced noise, and one test shows that performing noise reduction gave a worse result. The two significant tests are both for the symmetric case, indicating it might be worth to consider whether reducing noise is worthwhile for such cases. In order to evaluate this, some statistical information about the generated data was calculated, and is presented in Table 2.

Variable	Ref tree type:	Symmetric			Asymmetric		
	Mutation rate:	0.5	1.0	2.0	0.5	1.0	2.0
D	μ_D	0.1933	0.1233	0.2533	0.02667	-0.03667	-0.07333
	σ_D	1.367	1.652	2.289	0.7264	0.6191	0.6852
	standard error	0.0789	0.0954	0.132	0.0419	0.0357	0.0396
	95% conf. interval	0.193 ± 0.155	0.123 ± 0.187	0.253 ± 0.259	0.0267 ± 0.0822	-0.0367 ± 0.0701	-0.0733 ± 0.0775
D_noisy	μ_{noisy}	10.54	11.53	12.34	19.38	19.25	18.89
	σ_{noisy}	2.544	2.481	2.590	3.481	3.323	3.163
D_denoised	$\mu_{denoised}$	10.35	11.41	12.09	19.35	19.29	18.96
	$\sigma_{denoised}$	2.566	2.510	2.821	3.572	3.304	3.141

Table 2 - Statistical information on tree distances

The variable D_noisy represents the distance of the noisy tree and the reference tree, D_denoised the distance between the tree inferred from the noise reduced alignment and the reference tree, while D represents the difference D_noisy – D_denoised.

It is perhaps more interesting to look at the relative improvement of performing noise reduction. This is presented in Table 3 and is calculated as the ratio between μ_D and μ_{noisy}.

Ref tree type:	Symmetric			Asymmetric		
Mutation rate:	0.5	1.0	2.0	0.5	1.0	2.0
Relative improvement	1.83%	1.07%	2.05%	0.138%	-0.190%	-0.388%

Table 3 - Relative improvement obtained from performing noise reduction

As can be seen, even in cases which were shown significant, the largest increase is 2.05%, thus reinforcing the conclusion that noise reduction is not worthwhile.

Discussion

An experiment was performed on six data sets to evaluate whether performing noise reduction on protein mutlialignments is a worthwhile effort. The performance measure used was how closely the inferred tree from the noise reduced alignment resembles the reference tree when compared to the case before noise reduction. Out of six datasets, only two showed a statistically significant improvement, while one showed a worse result. Even for the two significant cases, the largest relative improvement was 2.05%, thus leading to the conclusion that noise reduction might not be worthwhile.

To compare the inferred trees to the reference tree, a Robinson-Foulds metric was used. One could also compare the number of times an inferred tree is equal to the reference tree. However, in most cases the two trees are different from the reference tree, but one of those trees might still resemble the reference tree more closely, so a tree comparison metric is needed.

Further investigation could be done with different distance metrics, to see whether the results would be the same. It could also be worth to compare how individual noise reduction rules affect the performance. Perhaps some of the rules are better left out.

Materials and Methods

The generated data consists of protein multialignments generated by the program `muscle` on sequences obtained by artificial evolution along the reference tree. For each such multialignment, another alignment was created by removing “bad” columns. The following rules were used to identify noisy columns:

- the column contains more than 50% indels
- at least 50% amino acids are unique
- no amino acid appears more than twice

To calculate the distance matrix from an alignment, the Jones-Taylor-Thornton substitution model was used, as implemented by the `phylip` program `protdist`. From the distance matrix, a tree was inferred using the neighbor-joining algorithm, as implemented by the `phylip` program `neighbor`. The distance between the inferred tree and the reference tree was then calculated by using the Robinson-Foulds metric, as implemented by the `DendroPy`’s `symmetric_difference` method.

This was done for each alignment in the data set, and a file was generated where every row contained the distance between the reference and the noisy tree, and the distance between the reference and the noise reduced tree.

This data was then statistically analyzed. The Wilcoxon's signed-rank test was used to evaluate significance of the improvement at the 5% level. The t-test showed identical results in regards to significance; therefore testing the data for normality was avoided.