

Automatic Recognition of Regions of Intrinsically Poor Multiple Alignment Using Machine Learning

Yunfeng Shan and Evangelos E. Milios

Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada B3H 1W5

Andrew J. Roger and Christian Blouin

Dept. of Biochemistry and Molecular Biology, Genome Atlantic/Genome Canada

Dalhousie University, Halifax, NS, Canada, B3H 1X5

Edward Susko

Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada, B3H 4H7

Abstract

Phylogenetic analysis requires alignment of gene or protein sequences. Some regions of genes evolve fast and suffer numerous insertion and deletion events and cannot be aligned reliably with automatic alignment algorithms. Such regions of intrinsically uncertain alignment are currently detected and deleted manually before performing phylogenetic analysis. We present the results of a machine learning approach to detect regions of poor alignment automatically. We compare the results obtained from Naive Bayes (NB), C4.5 Decision Tree (C4.5) and Support Vector Machine (SVM) approaches.

1 Introduction

In phylogenetic analysis, it is often necessary to edit multiple alignments in regions that are of intrinsically uncertain alignment. A manual approach is time-consuming for single genes and impossible for the large numbers of multiple sequence alignments generated in analysis of entire genomes. In this study, we examine the feasibility of using supervised machine learning techniques for detecting regions of poor alignment automatically.

2 Attribute Selection

The following three attributes were selected.

A. **Gap ratio** (g):

$$g = \frac{c}{t},$$

where c = number of gap characters, and t = number of taxa.

B. **Normalized site log likelihood ratio (NSLR)** (h):

$$h = \frac{\log(l) - \log(r)}{(1 - g) * t},$$

where l = site log likelihood given a preliminary tree, r = the site log likelihood if the sequences were unrelated (i.e. independent), g = gap ratio, and t = number of taxa.

C. **Consistency index** (CI):

$$CI = \frac{1}{PC},$$

where PC = parsimony count of gap to no-gap transitions given a preliminary tree.

Phylogenetic trees used to obtain the NSLR and the consistency index were calculated using distance matrices estimated by the TREE-PUZZLE 5.0 program under a Jones Taylor Thornton (JTT) plus gamma model followed by the neighbor-joining algorithm.

3 Training Data Collection

To generate a training set, 17821 sites from multiple sequence alignments were manually annotated by two phylogenetics experts to yield three site quality classes: bad, ambiguous and good.

A *random subset* was formed by randomly selecting a block of 100 sites from the training data without adjustment of the proportion of instances of the three classes in the subset. A *balanced subset* was similarly sampled with equal proportion of instances of the three classes. The entire data set and the random subset have the natural class distribution [1].

4 Experimental Design

We trained Naive Bayes (NB), C4.5 Decision Tree (C4.5) and Support Vector Machine (SVM) to recognize the three site quality classes. Performance was tested by using a two-way cross-validated experiment with default parameters [2]. This procedure was then repeated five times, each time using a different random seed for 10-fold cross-validation. For the SVM, three binary classifiers were built, each distinguishing between class K from the others.

5 Performance Measures

Three performance measures: precision, recall and correct rate were used [2]. *Precision* for class K is the proportion of items classified as K that are correct. *Recall* for class K is the proportion of items belonging to class K in the data set that were classified. *Correct rate* is the proportion of items that the classifier classified correctly for all classes.

6 Results and Discussion

6.1 Natural vs. balanced class distributions

Generally, the balanced subset was found to be better than the random subset or the whole dataset for NB and C4.5 (Table 1). This is consistent with other reports [1].

For the bad sites, SVM classifiers performed better if trained on the balanced subset rather than the random subset. SVM did not classify any sites as ambiguous and good sites. SVM classifiers trained on the balanced subset did not perform better than the random subset (Table 1).

Type of Data set	Predicted Class		
	Bad	Ambiguous	Good
a. NB:			
Entire data set	90.9/88.4	35.6/7.0	78.8/97.1
Random subset	89.6/91.4	NaN/0	59.3/98.3
Balance subset	96.6/95.0	84.4/20.4	55.4/98.3
b. C4.5:			
Entire data set	92.3/91.2	59.9/15.3	80.7/98.0
Random subset	90.3/89.3	64.2/16.2	63.1/95.2
Balance subset	93.8/94.2	64.7/87.1	87.2/56.7
c. SVM:			
Random subset	91.0/85.0	NaN/0	59.0/98.0
Balance subset	97.0/95.0	NaN/0	NaN/0

Table 1. Precision/Recall of three kinds of classifiers. Notation: NaN = all instances were classified into the “other” class(es), leading to division by zero.

6.2 Performances of three kinds of classifiers: NB vs. C4.5 vs. SVM

Based on the correct rate, C4.5 was better than NB (Table 1 and 2). Based on precision, C4.5 was better than NB based on the classifiers trained on the balanced subset. Based on recalls, there was no simple consistent pattern. For the bad sites, SVM was generally better than C4.5 and NB based on the correct rate and precision criteria, but not for the ambiguous and good sites (Table 1 and 2). More experiments are necessary to determine the optimal SVM parameter settings.

	a. NB	b. C4.5	c. SVM
Type of Data set			
Entire data set	80.4	91.2	NA
Random subset	67.8	89.3	77.0
Balance subset	71.1	94.2	78.2

Table 2. Correct classification rate of three kinds of classifiers. NA = Not Available.

7 Conclusions

We have demonstrated that supervised machine learning algorithms have the potential of accurately recognizing bad site in multiple sequence alignment based on attributes such as the gap ratio, consistency index and the NSLR. Among the techniques examined, NB and SVM classifiers are the best for bad site predictions, but C4.5 provides the best performance for the ambiguous and good site predictions.

It appears that the most difficult problem is to distinguish the ambiguous sites from the good sites, while distinguishing the bad sites is the easiest.

Generally, the classifiers of NB and C4.5 trained on the balanced subset achieve better classification performance compared with the natural class distributions.

Acknowledgements

We appreciate some of code written by Davin J. Butt. This work was supported in part by Grant #227085-00 from the NSERC, Canada awarded to AJR and is also part of a Genome Canada/Genome Atlantic large-scale project.

References

- [1] G. M. Weiss and F. Provost. *The effect of class distribution on classifier learning*. Technical Report ML-TR 43, Department of Computer Science, Rutgers University, 2001.
- [2] I. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques with java implementations*. 1999.