Gap Profiling: Scoring Indels in Multiple Sequence Alignment

Cory L. Strope  
Computer Science and Engineering  
University of Nebraska  
Lincoln, Nebraska 68588–0115  
Email: corystrope@gmail.com

Stephen D. Scott  
Computer Science and Engineering  
University of Nebraska  
Lincoln, Nebraska 68588–0115  
Email: sscott@cse.unl.edu

Etsuko N. Moriyama  
School of Biological Sciences,  
Center for Plant Science Innovation  
University of Nebraska  
Lincoln, Nebraska 68588–0660  
Email: emoriyama2@unlnotes.unl.edu

I. INTRODUCTION

Multiple sequence alignment (MSA) is most often the first step of bioinformatics and molecular evolutionary analyses. MSA is done by aligning the nucleotides or amino acids, hereafter referred to as characters, in columns based on their inferred homology by similarity, inserting gaps between positions in order to align the characters. Gaps represent insertion or deletion of portions of the sequence, or indels. In addition to characters, homology of the indels can also provide important information to molecular evolutionary analyses. However, gap information is usually ignored in many analyses.

Objective functions used in alignment methods often try to minimize the number of gaps by penalizing steps that introduce gap positions. The effectiveness of such objective functions to correctly place indels has been studied in only two unrealistic test cases (shown in [1]). These analyses were empirical and done in a case-by-case basis. Recent innovation in MSA methodology is to include homology information of indels into their objective functions [2], [3]. While these methods improved the character matching, gap placement accuracy is unknown. There is an absence of an indel representation that can be used to test indel placement accuracy.

Representing indels requires encoding the indels so that indels can be dealt with an equivalent manner as character data. Such indel coding methods have been used in, e.g., phylogenetic reconstruction [4]. Current methods, however, have drastic trade-offs between the number of indels used as information versus the number of falsely homologous indels inferred Due to this limitation, “indel coding” as been largely unused.

In this study, we introduce a novel method of representing indels, called gap profiling, and test the accuracy of inferring the placement of indels for the objective functions of seven MSA methods. We calculated sensitivity and specificity of indel placement in each method, and use a novel ROC-like measure to rate their accuracies. To assess the validity of our method, we compare the performance of our measure to the standard character-based Sum of Pairs scoring measure.

II. METHODS

Gap profiling is a novel technique for analyzing the accuracy of indel placement by comparing indel positions between two MSAs (i.e., a true or benchmark MSA vs an MSA for testing). We first break each MSA $M$ into the set of $N \times (N-1)$ pairwise alignments, where $N$ is the number of sequences in the multiple sequence alignment. For each pairwise alignment $(S_i, S_j)$, we remove columns that contain only gap characters, which are artifacts from the full MSA. We calculate the gap profile for $S_i$ against $S_j$, $GPs_{S_i,S_j}$, by incrementing the gap profile position between two characters in $S_i$ when there is one or more gap characters appearing between the two characters in $S_i$, as in Figure 1. This results in a binary gap profile the length of $S_i + 1$ sites long. After calculating the gap profiles of all pairwise alignments, we obtain a set of $N \times (N-1)$ gap profiles for $M$.

Using our representation, we define gap placement sensitivity as the distance of each gap in the true MSA, $GPs_{S_i,S_j}$, to the closest inferred gap in the reconstructed MSA, $GPs'_{S_i,S_j}$, and the gap placement specificity as the distance of each inferred gap in the $GPs'_{S_i,S_j}$ to the closest gap in $GPs_{S_i,S_j}$. We bin these distances as shown in Figure 2, which represent the distances from zero (exact match) to $w$, where $w$ is a given maximum value ($w = 4$ is used in Figure 2).
To test our indel representation, using indel-Seq-Gen version 2.0 [5], we simulated 20 datasets of 16 sequences, each 1000 characters long, varying the substitution and indel rates, number of taxa, and indel length distributions. The “true” MSAs were obtained from the simulation. The reconstructed MSAs were obtained by ClustalW2 [6], MAFFT [7], Muscle [8], ProbCons [9], PRANK [3] (using both PRANKs guide tree, and supplying the true guide tree), and FSA [2] (using standard and maximum sensitivity). Gap profiling was performed to compare these reconstructed and “true” MSAs.

We introduced a ROC-like sensitivity vs. specificity plot to assess method performance. We used a sliding decision line along the sensitivity bins (see Figure 2), starting at bin 0 to bin w, assuming everything to the left of the line is a true positive inference, while the rest are false positives. For each bin, we calculated the sensitivity as (TP)/(TP + FP). We use the same procedure in reverse (from bin w to bin 0, true negatives are on the right of the line) for specificity bins, where the specificity as (TN)/(TN + FN). We repeated this for w = 5, 10, 25, 50, 100, 250, 500, and 1000.

We measured the efficacy of the MSA reconstruction methods by the area under the curve (AUC) of this plot. We compare our measure of indel placement performance for each of the multiple sequence alignment methods against the standard performance measure that is based on character placement, the Sum of Pairs Score.

### III. RESULTS AND DISCUSSION

Results for the Gap Profile AUC are shown in Figure 3. The Gap Profile AUC generally correlates well with the Sum of Pairs score, although the Sum of Pairs score generally programs much higher. Since the substitution rate is low, the Sum of Pairs score is generally unaffected by the rate of indels. The possible exception is PRANK. PRANK is the only method that infers insertions and deletions as different events; inferring a deletion as an insertion will cause one or two character shifts in the alignment, which causes the Sum of Pairs score to be artificially low. The gap profile AUC, however, is robust to small shifts, while also penalizing methods that perform poorly at placing gaps (the so-called “gap magnets”, in which methods concatenate gaps in order to maximize their objective functions, which are based on maximizing the Sum of Pairs score), particularly ClustalW2. Both the Sum of Pairs score and the Gap Profile AUC show a declining trend as the proportion of insertions increases, since insertions cause MSAs to be much gappier, and methods create more gap magnet regions. Surprisingly, PRANK, which treats insertions and deletions differently, does not increase its Gap Profile AUC. More investigation is necessary to determine whether this is a problem with PRANK or with Gap Profile AUC. It may be necessary to consider a normalization step to account for the overrepresentation of insertions in MSAs. The gap profile AUC shows promise as a method for evaluating the accuracy of MSA methods in gap placement, leading to new ways to represent indels in objective functions used in MSA methods.

### IV. ACKNOWLEDGEMENT

This work was supported by NSF AToL grant 0732863.

### REFERENCES


