O alinhamento múltiplo de sequências é uma importante ferramenta de alicerce à pesquisa biológica, podendo ser usado em diversos âmbitos, por exemplo, identificação de estrutura secundária de uma proteína, detecção de RNA não codificante e inferência filogenética. Este último campo especificamente pode ser dramaticamente enviesado por alinhamentos pouco precisos, inclusive a filogenia resultante pode ser mais influenciada por alinhamentos que pelos métodos utilizados para a reconstrução filogenética (OGDEN & ROSENBERG, 2006). Citando Hall (2005), "*It is a truism that the quality of a tree is no better than the quality of the alignment used to estimate that tree*". Diante disto, pretende-se discorrer sobre métodos de alinhamento múltiplo amplamente utilizados com ClustalW e Muscle e também métodos que utilizam perfis de HMM, baseados em sequências intermediárias e informação sobre estrutura secundária, que são incorporados durante a busca do melhor alinhamento (LU & SZE, 2008) à luz das consequências para a reconstrução de filogenias.

Abstracts

Syst Biol. 2006 Apr;55(2):314-28.

Multiple sequence alignment accuracy and phylogenetic inference.

Ogdenw TH, Rosenberg MS.

Center for Evolutionary Functional Genomics, The Biodesign Institute, and the School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4501, USA. heath_ogden@asu.edu

Abstract

Phylogenies are often thought to be more dependent upon the specifics of the sequence alignment rather than on the method of reconstruction. Simulation of sequences containing insertion and deletion events was performed in order to determine the role that alignment accuracy plays during phylogenetic inference. Data sets were simulated for pectinate, balanced, and random tree shapes under different conditions (ultrametric equal branch length, ultrametric random branch length, nonultrametric random branch length). Comparisons between hypothesized alignments and true alignments enabled determination of two measures of alignment accuracy, that of the total data set and that of individual branches. In general, our results indicate that as alignment error increases, topological accuracy decreases. This trend was much more pronounced for data sets derived from more pectinate topologies. In contrast, for balanced, ultrametric, equal branch length tree shapes, alignment inaccuracy had little average effect on tree reconstruction. These conclusions are based on average trends of many analyses under different conditions, and any one specific analysis, independent of the alignment accuracy, may recover very accurate or inaccurate topologies. Maximum likelihood and Bayesian, in general, outperformed neighbor joining and maximum parsimony in terms of tree reconstruction accuracy. Results also indicated that as the length of the branch and of the neighboring branches increase, alignment accuracy decreases, and the length of the neighboring branches is the major factor in topological accuracy. Thus, multiple-sequence alignment can be an important factor in downstream effects on topological reconstruction.

J Comput Biol. 2008 Sep;15(7):767-77.

Multiple sequence alignment based on profile alignment of intermediate sequences.

Lu Y, Sze SH.

Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas, USA.

Abstract

Despite considerable efforts, it remains difficult to obtain accurate multiple sequence alignments. By using additional hits from database search of the input sequences, a few strategies have been proposed to significantly improve alignment accuracy, including the construction of profiles from the hits while performing profile alignment, the inclusion of high scoring hits into the input sequences, the use of intermediate sequence search to link distant homologs, and the use of secondary structure information. We develop an algorithm that integrates these strategies to further improve alignment accuracy by modifying the pair-Hidden Markov Model (HMM) approach in ProbCons to incorporate profiles of intermediate sequences from database search and utilize secondary structure predictions as in SPEM. We test our algorithm on a few sets of benchmark multiple alignments, including BAliBASE, HOMSTRAD, PREFAB, and SABmark, and show that it significantly outperforms MAFFT and ProbCons, which are among the best multiple alignment algorithms that do not utilize additional information, and SPEM, which is among the best multiple alignment algorithms that utilize additional hits from database search. The improvement in accuracy over SPEM can be as much as 5-10% when aligning divergent sequences. A software program that implements this approach (ISPAlign) is available at http://faculty.cs.tamu.edu/shsze/ispalign.

Despite considerable efforts, it remains difficult to obtain accurate multiple sequence alignments. By using additional hits from database search of the input sequences, a few strategies have been proposed to significantly improve alignment accuracy, including the construction of profiles from the hits while performing profile alignment, the inclusion of high scoring hits into the input sequences, the use of intermediate sequence search to link distant homologs, and the use of secondary structure information. We develop an algorithm that integrates these strategies to further improve alignment accuracy by modifying the pair–Hidden Markov Model (HMM) approach in ProbCons to incorporate profiles of intermediate sequences from database search and utilize secondary structure predictions as in SPEM. We test our algorithm on a few sets of benchmark multiple alignments, including BAliBASE, HOMSTRAD, PREFAB, and SABmark, and show that it significantly outperforms MAFFT and ProbCons, which are among the best multiple alignment algorithms that do not utilize additional information, and SPEM, which is among the best multiple alignment algorithms that utilize additional hits from database search. The improvement in accuracy over SPEM

can be as much as 5–10% when aligning divergent sequences. A software program that implements this approach (ISPAlign) is available at http://faculty.cs.tamu.edu/shsze/ispalign.