ngPhylo - N-Gram Modeled Proteins with Substitution Matrices for Phylogenetic Analysis

Brigitte Hofmeister
Department of Computer Science
Bucknell University
Lewisburg, PA 17837
bth007@bucknell.edu

Brian R. King *
Department of Computer Science
Bucknell University
Lewisburg, PA 17837
brian.king@bucknell.edu

ABSTRACT

Phylogenetic tree constructions are important for understanding evolution and species relatedness. Most methods require a multiple sequence alignment (MSA) to be performed prior to inducing the phylogenetic tree. MSAs, however, are computationally expensive and increasingly error prone as the number of sequences increase, as the average sequence length increases, and as the sequences in the set become more divergent. We introduce a new method called *ngPhylo*, an n-gram based method that addresses many of the limitations of MSA-based phylogenetic methods, and computes alignment-free phylogenetic analyses on large sets of proteins that also have long sequences. Unlike other methods, we incorporate the use of standard substitution matrices to improve similarity measures between sequences. Our results show that highly similar phylogenies are produced to existing MSA-based methods with less computational resources required.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and Genetics*

General Terms

Algorithms

Keywords

sequence analysis, phylogeny, n-gram model

1. INTRODUCTION

Phylogenetic analyses, specifically phylogenetic tree constructions, are important for understanding evolution and species relatedness. Currently, there are two main categories for constructing phylogenetic trees. Character-based methods, such as maximum parsimony [4] and maximum likelihood [3], require sequences to be of equal length, and therefore usually require a multiple sequence alignment (MSA) to be completed prior to analysis. Distance-based methods, such as UPGMA and neighbor joining methods, require a pair-wise distance matrix to be computed between all pairs of sequences. Like character-based methods, an MSA is used to infer distances between sequences.

Though performing an MSA is the standard first step for phylogenetic analysis, they have restrictions. The computational resources required to run an MSA is dependent on the number and length of the sequences aligned. Moreover, the accuracy of the alignment decreases in proportion to the number of sequences [8].

Alignment-free techniques have been employed, most which capitalize on analyzing distributions of fixed-length subsequences called n-grams. First proposed by Blaisdell in 1986 [1], the frequency and/or entropy of the n-grams have been used as feature vectors to compute a distance. Once the distances have been computed, the phylogenetic tree can be constructed using one of the distance based methods mentioned above.

Our method is a n-gram based approach that uses a specified substitution matrix to compute a biologically relevant measure of similarity between n-grams. A substitution matrix allows for matching of biologically similar n-grams. This provides a more meaningful phylogenetic distance calculation between sequences, and thus addresses limitations of strict n-gram matching used by existing alignment-free methods.

2. METHODS AND MATERIALS

2.1 N-Gram Model of a Protein

Each analyzed protein sequence, S, is converted to a frequency vector, $freq_S$ where each index in the vector represents the count of a specific n-gram occurring in S. An n-gram is defined as an overlapping subsequence of length n from an amino acid sequence. The frequency vector, $freq_S$, includes the count for all possible amino acid combinations of size n. Therefore, $|freq_S| = 20^n$, and $\sum_{i=1}^{20^n} freq_S[i] = length(S) - n + 1$.

2.2 Distance Computation and Tree Construction

The distance M between two arbitrary n-grams, denoted

*Corresponding author

BCB 13, September 22 - 25, 2013, Washington, DC
ACM 978-1-4503-2434-2/13/09.

w_a and w_b, is computed as follows:

$$M(w_a, w_b) = \sum_{i=1}^{n} (score(w_a[i], w_b[i])$$

where $score(w_a[i], w_b[i])$ is the substitution matrix value between the amino acids at index i in the n-grams being compared.

A standardized similarity score Sim between sequences S_i and S_j is computed in two phases. In phase one, we process all identical n-grams from S_i and S_j, decrement matched n-grams in $freq_{S_i}$ and $freq_{S_j}$ and update Sim accordingly. In phase two, we consider the remaining non-identical n-gram matches between S_i and S_j. We create a bipartite graph, where one set of nodes represent the remaining n-grams in S_i, denoted as $w_{i1}, ...w_{it}$; the nodes in the other set are those from S_j, denoted as $w_{j1}, ...w_{jv}$. For all pairs of nodes, an undirected, weighted edge, $E(w_{im}, w_{jp})$ is created if and only if $M(w_{im}, w_{jp}) > 0$. Edges are ordered by descending weight. The heaviest edge, $E(w_{i*}, w_{j*})$ is removed, its weight is added to $Sim(S_i, S_j)$, and nodes w_{i*} and w_{j*} are removed from each set (implying that all edges connected to w_{i*} and w_{j*} also are removed.) Edges are processed in this fashion until there are no more edges in the graph.

Finally, we standardize all Sim calculations to represent a distance:

$$dist(S_i, S_j) = 1 - \frac{Sim(S_i, S_j) - \min Sim(S_i, S_j)}{\max Sim(S_i, S_j) - \min Sim(S_i, S_j)}$$

The result is a complete distance matrix for all pairs of protein sequences. The phylogenetic trees were induced from the distance matrix using hierarchical clustering with single linkage updating.

2.3 Data

Three protein datasets were used for this study. D_1, a dataset of long, divergent proteins, is composed of 13 G-protein coupled receptor 98 proteins from a variety of animals. Each is over 6000 amino acids. D_2, a dataset of relatively short, divergent sequences, contains 21 serum albumin proteins that are about 630 amino acids. D_3, a large dataset of divergent proteins, consists of 200 prokaryote DNA gyrase subunit A sequences, each of which is about 875 amino acids.

3. RESULTS

We compared our method against ClustalW2 [5], Clustal Omega [7], and MUSCLE [2] using the default parameters. For our method, we used $n = 3$ for the n-gram length. This value was chosen for computational efficiency, and because we observed very minor improvement in tree quality for higher values of n. All methods were run three times under identical load conditions on the same system to reduce the variability due to the operating system.

Our results show that *ngPhylo* performed quite well against the MSAs. We also created a random tree with identical labels The phylogenetic trees produced by each method were compared against each other using the perl program TOPD [6] with the nodal analysis method [9] (Table 1).

The trees produced with *ngPhylo* compare well with those from ClustalW2, Clustal Omega, and MUSCLE. This is true for all datasets. *ngPhlyo* showed significant improvement in time for generating phylogenetic trees from D_1. However, it showed no improvement in time for D_2 and D_3.

Table 1: Nodal Analysis for Dataset D_1

	ngPhylo	Clustal Ω	ClustalW2	MUSCLE
ngPhylo				
Clustal Ω	1.144			
ClustalW2	0.716	1.240		
MUSCLE	0.716	1.240	0.000	
Random	2.684	2.833	2.455	2.455

4. DISCUSSION

Unlike many MSAs, our method is not limited by the number and length of sequences being analyzed. Also, our method would be able to applied to proteomes. This is not possible to do with MSAs because you cannot align a collection of proteins for one organism.

The *ngPhylo* method is currently implemented in Java, thus inherently slower than others written in C++. Even with this difference in speed due to programming language, the computational speed-up is encouraging. We are in the process of converting and optimizing the algorithm in C++. Also, we want to compare the results of our method on proteomes against existing studies. Finally, we plan to use the n-gram frequency and their index positions to determine which sections of a protein are less conserved.

5. REFERENCES

[1] B. E. Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 83(14):5155–9, July 1986.

[2] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–7, Jan. 2004.

[3] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–76, Jan. 1981.

[4] W. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20:406–416, 1971.

[5] e. a. Larkin, M A. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–8, Nov. 2007.

[6] P. Puigbò, S. Garcia-Vallvé, and J. O. McInerney. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics (Oxford, England)*, 23(12):1556–8, June 2007.

[7] e. a. Sievers, Fabian. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7:539, Jan. 2011.

[8] F. Sievers, D. Dineen, A. Wilm, and D. G. Higgins. Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics (Oxford, England)*, 29(8):989–95, Apr. 2013.

[9] M. A. Steel, D. Penny, and S. Url. Distributions of Tree Comparison Metrics-Some New Results. *Systematic Biology*, 42(2):126–141, 1993.