Whole Genome Phylogeny Based on Clustered Signature String Composition

Xiaomeng Wu, Guohui Lin*
Department of Computing Science
University of Alberta
Edmonton, Alberta T6G 2E8, Canada
xiaomeng,ghlin@cs.ualberta.ca

Xiu-Feng Wan, Dong Xu
Department of Computer Science
University of Missouri – Columbia
Columbia, Missouri 65211, USA
wanx,xudong@missouri.edu

Abstract

Peptide compositions constructed out of whole sets of protein sequences can be used as species signatures for phylogenetic analysis. To account for point mutations, an amino acid substitution model is integrated into the complete composition vectors through a novel peptide clustering algorithm. Such a refined signature is expected to highlight deeper evolutionary relationships among the species and employed into the whole genome phylogenetic analysis to define a new evolutionary distance measure. Computational experiments have been set up to validate the effectiveness of this new measure and a vertebrate evolutionary tree using a dataset of 832 *proteins for* 64 *vertebrates is reported.*

1. Introduction

The availability of an increasing number of completely sequenced genomes has opened up new avenues for understanding the evolution. In contrast to the traditional approaches where the molecular data is usually cautiously selected, the whole genomes afford unprecedented opportunities and perspectives for detecting evolutionary relationships at a micro point of view. However, this vast amount of sequence data challenge the phylogenetic analyses for evolutionary information representation to digest molecular sequences of millions of bytes.

The carefully selected data in the traditional approaches is relatively easy to be analyzed by adopting some substitution models that describe the prior knowledge about the evolution model. This seems to be advantageous over whole genomes in which gene transfer, unrecognized paralogy, and highly variable evolution rates exist. However, though the phylogenetic analysis by traditional approaches could provide accurate results on the selected molecular data, it doesn't tell well the species evolution since different sets of selected data normally result in conflicting phylogenetic analyses. On the other hand, whole genomes are believed to contain the complete evolutionary information and the phylogenetic analyses based on whole genomes are expected to equate the evolution of the organisms. Therefore, whole genome phylogeny becomes one of the major problems in comparative genomics [1]. The most profound difficulty in building phylogenies using whole genomes is to effectively and efficiently represent the evolutionary information hidden in whole genomes.

Traditional character-based phylogeny construction methods, including Maximum Parsimony (MP) and Maximum Likelihood (ML), build trees that optimize the distribution of the molecular data for each character, where substitution models are taken to align the multiple entries. Whole genomes for different organisms might contain different sets of genes in different sequential orders on the chromosomes. Therefore, multiple alignment can no longer be applied, not to mention its high computational complexity. During the past a few years, a considerable amount of efforts have been devoted to whole genome phylogeny study. All these efforts successfully avoid the high complexity stage of multiple alignment, and try to use the evolutionary information hidden in whole genomes as much as possible without using a substitution model. The main difference among these efforts is how they treat whole genomes to define the relative evolutionary distance between two whole genomes. Once the pairwise evolutionary distance matrix for the set of taxa is computed, they subsequently call distance-based phylogeny construction methods such as Neighbor-Joining [4] to build the tree.

Among several approaches, one category of whole genome phylogeny methods use the frequencies of segments of amino acids (or nucleotides if DNA sequences) as the species signatures. For example, frequencies of segments of all possible lengths are included in the *complete information set* [2]; linear combinations of frequencies of tri/tetra-peptides using a singular value decomposition are

*To whom correspondence should be addressed.

used in [5]; frequencies of length-k segments with random mutations subtracted, for a fixed k, are employed in the *composition vector* [3]. In [6], we proposed to integrate the ideas in the above methods to define a *complete composition vector* (CCV) for a whole genome, which composes of frequencies of length-k peptides with random mutations subtracted, for $k \in [k_{\min}, k_{\max}]$, where the length lower bound $k_{\min}$ and upper bound $k_{\max}$ were empirically determined (to be 3 and 7, respectively). In this poster, we present some partial results on adopting a substitution model, BLOSUM62, into the CCV-based whole genome phylogeny construction.

2. CCV-Based Phylogeny Construction

We assume that $s(\alpha_1\alpha_2\ldots\alpha_k)$ denotes the difference between the observed frequency and the expected frequency of peptide $\alpha_1\alpha_2\ldots\alpha_k$ through a second order Markov model. The k-th composition vector V^k is one with its entries recording $s(\alpha_1\alpha_2\ldots\alpha_k)$ for all length-k peptides. The union of V^k for $k \in [3, 7]$ is the CCV, which contains in total $1,347,368,000$ entries. Needless to say, it contains many 0-entries and most of the others are insignificant, which means perhaps all CCVs would have the similar entry values. To reduce the dimensionality, we use matrix BLOSUM62 to classify amino acids into 15 groups and assign a code for each group. Subsequently, every peptide receives a code or a set of codes, which is used to cluster the peptides. Essentially, peptides within a cluster form a *critical clique*, which tells that they are similar to each other and they are similar to a common set of peptides outside. CCVs are then refined to have the entries merged into one if the corresponding peptides are in a cluster.

These refined CCVs represent the species in a high dimensional space and the cosines of angles between them are used to measure the pairwise evolutionary distances. At the time a pairwise distance matrix for the species is calculated, Neighbor-Joining method [4] is employed to construct a phylogeny.

3. Experiments

We have tested the CCV-based phylogeny construction method on a number of datasets, one is included in this poster. The dataset is from [5] and it contains in total 832 mitochondrial proteins obtained from the whole mitochondrial genomes for 64 vertebrates, where every species has 13 homologous proteins. The resultant phylogeny is shown in Figure 1, where the numbers labeling the branches are bootstrapping results from 200 iterations. This output phylogeny maps well to the taxonomy tree, and it even smooths out disagreements occurring in the phylogeny by pure CCVs and the phylogeny through SVD method.

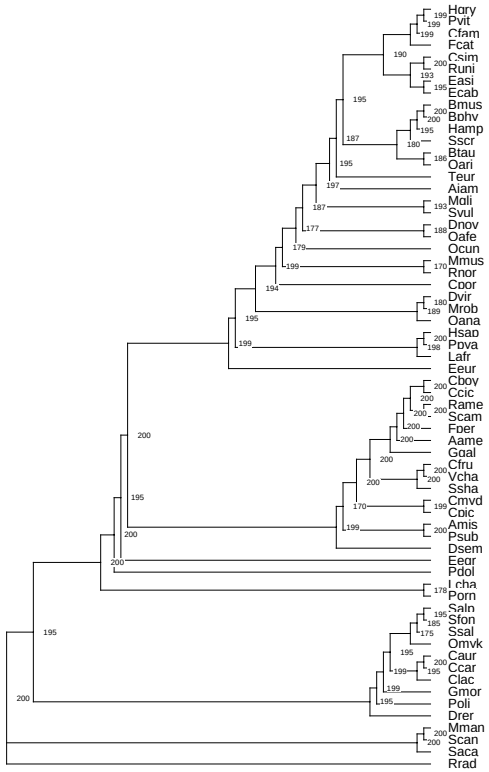

Figure 1. The consensus species tree on the 64 vertebrates based on refined CCVs.

References

[1] E. V. Koonin. The emerging paradigm and open problems in comparative genomics. *Bioinformatics*, 15:265–266, 1999.

[2] W. Li, W. Fang, L. Ling, J. Wang, Z. Xuan, and R. Chen. Phylogeny based on whole genome as inferred from complete information set analysis. *Journal of Biological Physics*, 28:439–447, 2002.

[3] J. Qi, B. Wang, and B.-L. Hao. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. *Journal of Molecular Evolution*, 58:1–11, 2004.

[4] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[5] G. Stuart, K. Moffet, and J. Leader. A comprehensive vertebrate phylogeny using vector representation of protein sequences from whole genomes. *Molecular Biology and Evolution*, 19:554–562, 2002.

[6] X. Wu, X. Wan, G. Wu, D. Xu, and G.-H. Lin. Whole genome phylogeny construction via complete composition vectors. Technical Report TR05-06, Department of Computing Science, University of Alberta, January 2005.