Information Entropy Based Methods for Genome Comparison

Mehul Jani
Department of Biological Sciences,
University of North Texas
Denton, Texas 76203
mehuljani@my.unt.edu

Rajeev K. Azad
Departments of Biological Sciences and
Department of Mathematics,
University of North Texas
Denton, Texas 76203
rajeev.azad@unt.edu

ABSTRACT

A plethora of biologically useful information lies obscured in the genomes of organisms. Encoded within the genome of an organism is the information about its evolutionary history. Evolutionary signals are scattered throughout the genome. Bioinformatics approaches are frequently invoked to deconstruct the evolutionary patterns underlying genomes, which are difficult to decipher using traditional laboratory experiments. However, interpreting constantly evolving genomes is a non-trivial task for bioinformaticians. Processes such as mutations, recombinations, insertions and deletions make genomes not only heterogeneous and difficult to decipher but also renders direct sequence comparison less effective. Here we present a brief overview of the sequence comparison methods with a focus on recently proposed alignment-free sequence comparison methods based on Shannon information entropy. Many of these sequence comparison methods have been adapted to construct phylogenetic trees to infer relationships among organisms.

General Terms

Algorithms, Measurement, Performance, Design, Reliability, Verification.

Keywords

Genome comparison, Sequence alignment, Shannon entropy, Segmentation, Clustering

1. INTRODUCTION

Evolution has always intrigued humans. Early approaches of studying evolution were based on fossil study and other palaeontological methods. Developments in the field of molecular biology, especially DNA sequencing, led to methods based on comparing DNA and protein sequences for studying the relatedness between the DNA sequence and thus the organisms. Phylogenetic trees constructed using simple alignment methods were used to infer evolutionary relationships among organisms. Phylogenetic trees are used in diverse fields such as systems biology, molecular biology, Darwinian medicine and ecology. Often such studies have helped address disease problems, e.g., phylogenetic trees of bacteria have advanced our understanding of the spread of antibiotic resistance patterns and the emergence of virulence.

With the advent of high throughput and next generation sequencing technologies, we now have access to complete genomes of over 4000 prokaryotes and over 180 eukaryotes (http://www.genomesonline.org). Such vast amount of genomic data calls for development of robust bioinformatics approaches to studying evolution via inferring phylogeny relationships among organisms. There are broadly two approaches used for comparing DNA sequences- alignment based sequence comparison and alignment-free sequence comparison, which we discuss briefly in the next sections.

2. ALIGNMENT BASED SEQUENCE COMPARISON

Sequence alignment is the most common approach used for comparing sequences. In pairwise sequence alignment, an optimal alignment between two given sequences is searched for by maximizing a scoring function which is essentially the sum of the residue to residue alignment scores between the sequences. Dynamic programming methods were developed for the pairwise global alignment and for the local alignment between sequences. The former, the Needleman-Wunsch algorithm [10], aligns two sequences end to end, that is, the full span of the sequences are aligned against each other, while the latter, the Smith-Waterman algorithm [16], searches for best aligned motif (short conserved regions) in the alignment. Heuristic approaches for local sequence alignment were developed later for fast and efficient alignment of long biological sequences. Among these, FASTA [11] and BLAST [1] are the most frequently used heuristic algorithms. The visualization techniques such as Dot matrix were developed for visualizing the alignment between two sequences; this is among the oldest approaches for sequence comparison, first used by Gibbs and McIntyre in 1970. Biological sequences come in family; homologous sequences in a family share the common ancestry and often have similar functions. Pairwise alignment methods are limited in their ability to detect members of a sequence family. To circumvent these limitations and to detect remote homologs, multiple sequence alignment methods were developed. The progressive alignment methods first align the most closely related pair of sequences and then the next most similar sequence to this pair is aligned and the process is repeated iteratively to build a multiple sequence alignment (also sometimes referred to as 'profile'). CLUSTALW [17] is one of the most popular tools used for multiple sequence alignment. Multiple sequence alignment is a precursor to phylogenetic tree construction. Based on the alignment score between sequences in a profile, a distance matrix is created and a phylogeny tree is constructed using the distance matrix. Evolutionary relationships are thus inferred from relative positions of sequences in a phylogenetic tree.

Genome wide sequence alignment is a huge computational burden. Often the organismal relationships are inferred by constructing trees using highly conserved nucleotide sequences of RNA genes or the conserved sequences of proteins. The problem with using only conserved gene or protein sequences is that the evolutionary signals from rest of the genome are ignored. Since evolutionary signals are dispersed throughout the genome and not just restricted to a few genes, ignoring these signals may have confounding implications. In fact, trees made using conserved RNA or protein sequences have been shown to contradict each other. Further, most alignment methods do not account for long range interactions within genomes. Moreover, natural evolutionary processes like recombinations, mutations, deletions, insertions, rearrangements, etc., make direct alignment between sequences difficult, especially when such changes happen frequently leading to fast evolving genomes with little evolutionary signals for a reliable sequence alignment. In general, sequence alignment method works best when the sequences being compared share high homology. Therefore, there is a great need of methods that can adequately account for evolutionary signals underlying the genomes of organisms.

3. ALIGNMENT-FREE SEQUENCE COMPARISON

Alignment-free approaches are especially useful if the sequences do not share high homology or are rapidly accumulating changes thus obfuscating the evolutionary signals. Frequent rearrangements, in particular, disrupt the sequence contiguity and thus render such sequences unalignable in order to assess their common ancestry. To circumvent the limitations of alignment based methods, several approaches that do not require alignment for sequence comparison have been proposed. These so called alignment-free methods are based on *k*-mer frequency for computing the similarity (or dissimilarity) score between the sequences. The goal of such methods is to assess the divergence between two sequences in terms of difference in the frequency distributions of *k*-mers in the sequences. The frequently used distance measures to assess the sequence divergence include Euclidean distance [2], *d2* distance [18], covariance or correlation function [12], Mahalanobis distance [21], Kullback-Leibler divergence [22] and Kolmogorov complexity metric [7]. In a different approach for alignment-free sequence comparison, methods based on substrings [5, 19] have been used. The average common substring (ACS) method by Ulitsky et al [19] calculates average length of maximum common substrings for every site of each sequence and then pairwise genome sequence distance is calculated [19]. B. Haubold et al [5] used the shortest unique substrings in a set of sequences being studied for sequence comparison. Recently, J. Cheng et al [6] have built a multi-methods web server for alignment-free genome phylogeny, which can implement 12 popular alignment-free methods in a user friendly web platform. We refer the readers to Vinga and Almeida [20] for a comprehensive review of some of the alignment free methods discussed above.

Sims et al [15] proposed a feature frequency profile (FFP) method, a method based on *k*-mer frequency approach, which was shown to outperform other methods including the average common substring and Gencompress [3] methods. In this method, the frequencies of all possible features (the *k*-mers) of size *k* are computed to make a feature frequency profile. The total number of possible features will be 4^k in DNA sequence comparison. The difference between two genomic sequences, quantified in terms of difference in their *k*-mer compositional biases, was computed using Shannon information entropy based measure (Eqn. 1 in Section 4). The most important contribution of this method, as noted by the authors, is obtaining the optimal *k*-mer size to be used for sequence comparison. The lower limit of the *k*-mer can be empirically obtained, whereas upper limit of *k*-mer is calculated based on cumulative relative entropy. In order to infer organismal relationships, the information-entropic measure, namely, the Jensen-Shannon divergence (Eqn. 1), was used to compute the distance between the genome sequences of organisms and then a phylogenetic tree was constructed using this distance matrix.

The performance of FFP method and other *k*-mer frequency based methods for alignment-free sequence comparison depends on the *k*-mer size [15]. While longer *k*-mers carry more information and therefore confer greater predictive power to the methods, it is, however, not practical to use longer *k*-mers if the sequences under comparison are not sufficiently long enough. In contrast, shorter *k*-mers provide reliable statistics, however, this may represent the inherent stochastic nature of genomes rather than having any biological or phylogenetic meaning.

Genomes are inherently heterogeneous. Bacterial genomes are chimeras of genes with different ancestry. Genomic mosaicism also arises when different segments of a genome are subject to different evolutionary pressures. All alignment-free methods including the FFP method represent a genome sequence as a *k*-mer frequency distribution, thus ignoring the inherent genomic mosaicism that requires multiple *k*-mer frequency distributions to represent uniquely distinct sequence classes within a mosaic genome. Methods that use a single oligomer distribution as the representation of a genome can yield confounding results when comparing two or more mosaic genomes. A single oligomer distribution averages out evolutionary signal from entire genome, disregarding the heterogeneity of the genome [13]. To overcome this problem, Azad and Li [13], first deconstructed the intragenomic heterogeneity using Shannon information entropy based recursive segmentation and clustering method, and then compared the compositionally homogenous regions from the genomes of interest.

4. RECURSIVE SEGMENTATION AND AGGLOMERATIVE CLUSTERING

An integrative framework of recursive segmentation and agglomerative clustering was developed recently to deconstruct the complex heterogeneities within genomic data [13]. Recursive segmentation for DNA sequence analysis has a history of over a decade [4,8]. The recursive segmentation and agglomerative clustering method interprets genomic data at the intrusive level of complexities using Shannon entropy [14]. This method uses Jensen-Shannon (JS) divergence measure for assessing divergence or dissimilarity between two sequences. The Jensen-Shannon divergence between two sequences S_1 and S_2 can be measured using the following formula [9],

$$D\ (S_1,S_2) = H(S) - \pi_1 H(S_1) - \pi_2 H(S_2). \tag{1}$$

Here, Shannon entropy H for a sequence is defined as $H = -\sum_x p(x) \log_2 p(x)$, where $p(x)$ is the probability of (oligo)nucleotide (residue for protein sequences) x estimated from the count of x in the sequence. S is the concatenation S_1 and S_2, and π_i is the weight factor proportional to the length of S_i, $\sum_i \pi_i = 1$. The entropy function measures the information stored in a sequence.

The genome complexity is decomposed successively by performing a binary segmentation recursively until none of the sequence segments or regions can be divided further [9] using following steps: (i) For a sequence S, the difference between sequence segments left and right to each sequence position in S is calculated using Jensen-Shannon divergence measure. (ii) The position of maximum divergence between the left and right sequence segments is located. (iii) The sequence is segmented at this position to get two segments, S_1 and S_2 provided the segmentation is deemed statistically significant. (iv)The aforementioned procedure is repeated for segments S_1 and S_2 recursively until none of the resulting sequence segments can be divided further. (v) These compositionally homogeneous sequence segments are now considered as distinct clusters, each segment assigned to a distinct cluster. In this step, similar contiguous segment clusters are identified and grouped together (vi) These segment clusters are the seed clusters for the next step of the clustering procedure. The grouping of similar clusters is followed recursively until the difference between any two clusters becomes significantly large. This last step clusters even non-contiguous segments and thus account for long range interactions or relationships between different regions in a genome.

This recursive segmentation procedure can be accomplished within a hypothesis-testing framework [4] or a model-selection framework [8]. Azad and Li [13] allowed hyper-segmentation in the hypothesis testing framework. This helped to increase the sensitivity of the method in identifying the break points or segment boundaries. However, hyper-segmentation may cause fragmentation of biologically important domains. To reestablish the segmental structure, segmentation was followed by clustering (step v above) at a relaxed clustering stringency.

To assess the divergence between genomes, Genome Wide Distance (GWD) was calculated using following formula:

$$GWD = \frac{1}{2}\left[\frac{1}{M}\sum_{i=1}^{M}\min\{D(G_1^i, G_2^1), \ldots, D(G_1^i, G_2^N)\} + \frac{1}{N}\sum_{j=1}^{N}\min\{D(G_1^1, G_2^j), \ldots, D(G_1^M, G_2^j)\}\right] \tag{2}$$

Here, $D(G^i{}_1, G^j{}_2)$ is the Jensen-Shannon divergence between clusters i and j of genomes G_1 and G_2. M and N are number of clusters for genome G_1 and G_2 respectively.
This method was reported to perform better than the FFP method for comparing genomes [13]. This validated the hypothesis that relationships among organisms could be better explained by first decomposing their genome complexities and then comparing compositionally distinct components of their genomes. In the recursive segmentation and agglomerative clustering approach, the global genomic heterogeneity is deciphered first; the earlier obtained split points thus guide the next rounds of segmentation to decipher the local heterogeneities and in this process, eventually, the distinct evolutionary signals encoded in biological domains within a genome are deciphered. This method thus captures the evolutionary patterns within genomes reflecting disparate evolutionary trajectories, thus helping in deducing the evolutionary relationships among organisms.

This method was used to address several other pressing issues in biology, such as, identification of alien genes in bacterial genomes and detection of copy number variations in cancer genomes [13]. In future, this method can be adapted to detect other biologically important features such as isochores or the origin and terminus of replication.

5. CONCLUSIONS

The vast number of methods developed for comparing genome sequences highlights the significance of deducing reliable phylogenetic relationships. Traditional sequence alignment methods though reliable for sequences which are highly related or share high homology often prove to be deficient when comparing rapidly evolving sequences.

Alignment-free approaches have made significant progress since it was first used by Blaisdell [2]. Many recent methods for sequence comparison have used alignment-free approach. The alignment-free approach allows computing distances between large genomes in relatively less time. Alignment-free methods are more robust for comparing highly evolved sequences, sequences which have undergone changes at multiple loci in a chromosome, and even shorter sequences.

The advantage of using recursive segmentation and agglomerative clustering method is that it first decomposes the complexities of heterogeneous genomes and then compares the homogeneous parts of the genomes, thus providing a better comparison tool for elucidating organismal relationships. This method can be used in concert with alignment based methods to construct robust phylogenetic trees.

6. ACKNOWLEDGEMENTS

This work was supported by a faculty start up fund from the University of North Texas to R.K.A.

7. REFERENCES

[1] Altschul SF, Gish W, Miller W., Myers EW and Lipman DJ 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

[2] Blaisdell BE 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl Acad. Sci. USA 83, 5155–5159.

[3] Chen X, Kwong S and Li M 2001. A compression algorithm for DNA sequences. IEEE Engineering in Medicine and Biology Magazine 20, 61–66.

[4] Grosse I, Bernaola-Galvan P, Carpena P, Roman-Roldan R, Oliver J, Stanley HE 2002. Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. Phys Rev E 65:041905.

[5] Haubold B, Pierstorff N, Möller F, Wiehe T 2005. Genome comparison without alignment using shortest unique substrings. BMC Bioinformatics 6:123.

[6] Cheng J, Cao F and Liu Z 2013. AGP: A multi-methods web server for alignment-free genome phylogeny. Mol. Biol. Evol. 30:1032–1037

[7] Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17:149–154.

[8] Li W 2001. New stopping criteria for segmenting DNA sequences. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 86:5815–5818.

[9] Lin J 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inform. Theory 37:145–151.

[10] Needleman SB and Wunsch CD 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

[11] Pearson WR 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183, 63–98.

[12] Petrilli P 1993. Classification of protein sequences by their dipeptide composition. Comput. Appl. Biosci. 9:205–209.

[13] Azad RK and Li J 2013. Interpreting genomic data via entropic dissection. Nucleic Acids Res. 41: e23.

[14] Shannon CE 1948. A mathematical theory of communication. The Bell System Technical J. 27: 379–423.

[15] Sims GE, Jun SR, Wu GA, Kim SH 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl Acad. Sci. USA. 106, 2677–2682.

[16] Smith TF and Waterman MS 1981. Identification of common molecular subsequences. J Mol Biol. 147:195-197.

[17] Thompson JD, Higgins DG, Gibson TJ 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

[18] Torney DC, Burkes C, Davidson D, Sirkin KM 1990. In: Computation of d2: A measure of sequence dissimilarity, computers and DNA, SFI studies in the sciences of complexity. Bell G, Marr T, editors. VII. New York, NY: Addison-Wesley.

[19] Ulitsky I, Burnstein D, Tuller T, Chor B 2006. The average common substring approach to phylogenomic reconstruction. Journal of Computational Biology 13, 336–350.

[20] Vinga S, Almeida J 2003. Alignment free sequence comparison- a review, Bioinformatics 19: 513-523.

[21] Wu TJ, Burke JP, Davison DB 1997. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53:1431-1439.

[22] Wu TJ, Hsieh YC, Li LA 2001. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics. 57:441–448.