Computational and Structural Biotechnology Journal 13 (2015) 352–357

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Homology-Independent Metrics for Comparative Genomics

Tarcisio José Domingos Coutinho [a], Glória Regina Franco [a], Francisco Pereira Lobo [b,*]

[a] Departamento de Bioquímica e Imunologia, Programa de pós-graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Antonio Carlos Avenue, 6627, Belo Horizonte, Minas Gerais CEP 31270-901, Brazil
[b] Laboratório Multiusuário de Bioinformática, Embrapa Informática Agropecuária, André Tosello Avenue, 209, Barão Geraldo, Campinas, São Paulo CEP 13083-886, Brazil

ARTICLE INFO

Article history:
Received 4 February 2015
Received in revised form 6 April 2015
Accepted 18 April 2015
Available online 5 May 2015

Keywords:
Comparative genomics
Homology-independent metrics
Genomic signatures

ABSTRACT

A mainstream procedure to analyze the wealth of genomic data available nowadays is the detection of homologous regions shared across genomes, followed by the extraction of biological information from the patterns of conservation and variation observed in such regions. Although of pivotal importance, comparative genomic procedures that rely on homology inference are obviously not applicable if no homologous regions are detectable. This fact excludes a considerable portion of "genomic dark matter" with no significant similarity — and, consequently, no inferred homology to any other known sequence — from several downstream comparative genomic methods. In this review we compile several sequence metrics that do not rely on homology inference and can be used to compare nucleotide sequences and extract biologically meaningful information from them. These metrics comprise several compositional parameters calculated from sequence data alone, such as GC content, dinucleotide odds ratio, and several codon bias metrics. They also share other interesting properties, such as pervasiveness (patterns persist on smaller scales) and phylogenetic signal. We also cite examples where these homology-independent metrics have been successfully applied to support several bioinformatics challenges, such as taxonomic classification of biological sequences without homology inference. They where also used to detect higher-order patterns of interactions in biological systems, ranging from detecting coevolutionary trends between the genomes of viruses and their hosts to characterization of gene pools of entire microbial communities. We argue that, if correctly understood and applied, homology-independent metrics can add important layers of biological information in comparative genomic studies without prior homology inference.

Contents

Abbreviations: DOR, dinucleotide odds ratio; GC3S, frequency of G + C at the third position of synonymous codons; HI, homology-independent; HD, homology-dependent; ORF, open reading frame; RSCU, relative synonymous codon usage; NC, effective number of codons; NC′, effective number of codons considering GC3S.

* Corresponding author. Tel.: +55 19 32115843.
E-mail addresses: coutinho.tarcisio@gmail.com (T.J.D. Coutinho), gfrancoufmg@gmail.com (G.R. Franco), franciscolobo@gmail.com, francisco.lobo@embrapa.br (F.P. Lobo).

http://dx.doi.org/10.1016/j.csbj.2015.04.005

1. Introduction

Most computational methods available for comparative genomics rely on initial similarity searches to infer homology relationships and, consequently, analyze the wealth of genomic data currently available. Among others, current methods for comparative genomic analysis based on the detection of homologous sequences allow the 1) determination of the time of divergence between taxa through the theory of molecular clock [1]; 2) automatic annotation of new genomes based on orthology inference [2]; 3) estimation of the rates of evolution of protein families [3]; 4) analysis of the overall evolution of genomes through genome-scale analysis of patterns of gain/loss of genomic elements [4]; 5) searching for higher-order layers of positional genomic information (haplotypic blocks, synteny, etc.) [5]; and 6) genomic-scale search for patterns of positive Darwinian selection [6], among many others.

The computational methods for comparative genomic analysis based in the detection of homologous regions, from now on referred as homology-dependent (HD) methods, although crucial for several bioinformatic pipelines, are limited to genomic sequences with detectable homologous regions, usually identified through computationally intensive software that contains somewhat arbitrary cut-offs to define groups of homologous sequences [7,8]. The failure to detect such regions prevents the application of virtually any HD method and excludes several interesting classes of DNA sequences from further analysis. ORFans — orphans Open Reading Frames (ORFs) without any detectable similarity to other sequences — are commonly found in complete genomes and in environmental sequences and constitute a true "dark matter" of biological data that cannot be surveyed using traditional HD methods [9,10]. In some newly discovered taxa, such as the large DNA viruses from *Mimiviridae*, the vast majority of coding sequences do not share significant similarity with known proteins [11].

In this mini-review we compile a list of metrics used in comparative genomic studies that share an unusual property for this purpose: they do not rely on initial homology inference, and can be calculated from individual sequence data alone. Such metrics, from now on referred as homology-independent (HI) metrics, can be easily calculated for virtually any fragment of any genome that fulfills a few criteria, such as minimum length and complexity. These metrics usually detect biases by comparing the observed frequencies of nucleotide words, especially dinucleotide and codons, with expected frequencies for the same words. The dinucleotide usage patterns in a given genome are commonly referred in the scientific literature as genomic signatures, since they are also taxon-specific and highly conserved in a given genome [12–14]. Here we also highlight the relative strengths and weaknesses of such metrics and report comparative genomic studies that applied such metrics to extract biologically meaningful information that would be otherwise impossible to obtain using common HD comparative genomic methods.

2. Homology-Independent Metrics: Causes and Properties

2.1. Causes of Variation in Homology-Independent Metrics

Most explanations for the biased values observed in HI metrics are due to a complex interplay between three broad groups of phenomena that shape together the use of nucleotide words in genomes. One of these groups is composed of mutational pressures where a given nucleotide word is significantly more (or less) used than its expected frequency due to mutational events. Possible sources of mutational pressures are distinct transition/transversion ratios [15], CpG underrepresentation in vertebrate genomes due to methylation/deamination processes occurring in this dinucleotide [16] and distinct nucleotide incorporation efficiency by polymerases during genome replication [17], among many others.

A second group of phenomena responsible for HI biases in genomes is composed of selection pressure events, in which natural selection shapes the differential usage of nucleotide motifs. In fact, several nucleotide motifs (such as the "TATA box") interact with the transcription/translation machinery and are classic examples of conservation of nucleotide words in genomic sequences due to selection pressure [18]. Another classic cause of variation in nucleotide words (codons) induced by selection pressure is the more-than-expected usage of synonymous codons that corresponds to the more abundant aminoacyl-tRNAs in cell cytoplasm in order to increase translation speed/efficiency, a selection pressure particularly strong in single-celled organisms [19] and in highly-expressed genes [20]. A final broad class of factors known to influence the use of nucleotide words in genomes is the occurrence of neutral processes such as genetic drift during the course of evolution [21]. Therefore, if properly modeled and interpreted, the results obtained through HI metrics in comparative genomic studies can highlight broad patterns of mutational and selective forces as well as random variations acting in the genomes under analysis.

2.2. General Properties of Homology-Independent Metrics

Besides not requiring previous homology relationships to analyze genomic data, HI metrics contain other general properties shared by most or all of them. An interesting property is that most HI metrics contain null models that take into account major factors already known to influence the frequencies of nucleotide words. Different HI metrics consider factors such as GC content, observed frequencies of smaller words that compose the word under analysis, degeneracy of the genetic code and amino acid usage, among others, when calculating null models. Therefore, any bias detected using HI metrics with proper null models is not explainable by factors already taken into account when computing null models, and represent biological phenomena that require further explanation/investigation. Also, several HI metrics are pervasive in the sense that values for whole genomes should persist at smaller scales. Some HI metrics remain reasonably constant for fragments with as few as 125 base pairs when compared with values calculated for entire genomes, or even when comparing coding and non-coding regions of genomes, making HI metrics a robust option to develop procedures to classify nucleotide sequences to taxonomic units, as in the case of genomic signatures [13].

Another useful aspect of HI metrics when applied to comparative genomics is the fact that some of them generate results that contain phylogenetic signal and are able to represent phylogenetic relationships, arguably with a more global view of the evolutionary process [22]. The patterns of DOR and codon usage bias in complete genomes of prokaryotes present a strong correlation with phylogenetic trees of 16S ribosomal RNA and housekeeping genes [14,23]. Comparative genomics using HI metrics may better reflect the global phylogenetic relationships between complete genomes by considering, for instance, events of horizontal transfer (HGT, one of the many factors known to change local frequencies of nucleotide words in genomic sequences) as part of the evolutionary signal in opposition to the reductionist analysis of single genes as proxies to faithfully represent the phylogenetic history of the entire genome.