Computational and Structural Biotechnology Journal 14 (2016) 298–303

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Well-characterized sequence features of eukaryote genomes and implications for *ab initio* gene prediction

Ying Huang [b], Shi-Yi Chen [a,*], Feilong Deng [a]

[a] Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu 611130, China
[b] College of Veterinary Medicine, Sichuan Agricultural University, Chengdu 611130, China

ARTICLE INFO

Article history:
Received 23 April 2016
Received in revised form 6 July 2016
Accepted 12 July 2016
Available online 27 July 2016

Keywords:
Sequence features
Compositional properties
Functional signals
Ab initio gene prediction
Eukaryotes

ABSTRACT

In silico analysis of DNA sequences is an important area of computational biology in the post-genomic era. Over the past two decades, computational approaches for *ab initio* prediction of gene structure from genome sequence alone have largely facilitated our understanding on a variety of biological questions. Although the computational prediction of protein-coding genes has already been well-established, we are also facing challenges to robustly find the non-coding RNA genes, such as miRNA and lncRNA. Two main aspects of *ab initio* gene prediction include the computed values for describing sequence features and used algorithm for training the discriminant function, and by which different combinations are employed into various bioinformatic tools. Herein, we briefly review these well-characterized sequence features in eukaryote genomes and applications to *ab initio* gene prediction. The main purpose of this article is to provide an overview to beginners who aim to develop the related bioinformatic tools.

Due to tremendous progresses in terms of efficiency, accuracy and cost for the high-throughput sequencing technologies, a large number of genome sequences of eukaryotic, prokaryotic and archaea organisms are increasingly becoming available [1,2]. These efforts are expected to open the window for better understanding the kinds of biological processes because essential information in principle is encoded in genome sequences. Nevertheless, it is also challenging for meaningfully decoding the huge amount of DNA sequences; for example, we are still infants in understanding biological implications of the substantial fraction sequences of "junk DNA" in eukaryote genomes, which don't encode any known proteins [3]. Additionally, a recent publication also revealed that the sequence context has functional consequences by influencing the substitution rate of adjacent nucleotides [4], which would complicate the biological explanation of genome sequences because the more complex mathematical models would be required.

By contrast to experimental investigations on biological functions, the *in silico* analysis of DNA sequences is essential in post-genomic era. There are many general properties of DNA sequence, such as GC content and base composition, having been well used for *in silico* analysis [5]. Additionally, *ab initio* prediction of gene structure is a critical step after sequencing whole genome and therefore has received much attention over the past decade [6]. Because of limitations of biological knowledge and bioinformatic algorithm, however, it still remains to be further improved on precision for these existing bioinformatic tools of gene prediction. In the present article, we briefly review these well-characterized features of DNA sequence and applications to *ab initio* gene prediction in eukaryotes. Although some literatures were published more than ten years ago, it is still helpful to provide an overall landscape for promoting the development of bioinformatic tools. Also, genome architectures for these available eukaryotic species are summarily illustrated in advance.

1. Outlines of genome architecture

To explore the evolutionary dynamics and biological consequences on genome size, base composition, and relative proportions of functional and nonfunctional sequences are deemed fascinating challenges in biology. The transposable genetic elements, in combination with natural selection, have been acknowledged to contribute to genome evolution, which result into considerable accumulation of repetitive sequences [7–9]. However, many proposed mechanisms trying to account for the genome evolution still remain uncertain or controversial, and these topics are also beyond scope of the present review. Fortunately, the recently prevailing approach of pan-genome analysis would be anticipated to provide more insights into this field [10].

According to intuitive expectation, the genome size would be proportional to species complexity, *i.e.*, the higher organisms have larger genomes. However, substantial variability of DNA content per haploid

* Corresponding author at: Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, 211# Huimin Road, Wenjiang 611130, Sichuan, China.
E-mail address: sychensau@gmail.com (S.-Y. Chen).

http://dx.doi.org/10.1016/j.csbj.2016.07.002

genome (C-value) have been widely observed even among the closely related species from same genus [3], which is thereby termed the C-value paradox. Scientific publications in eukaryotes on diversity patterns, evolutionary mechanisms and research methodologies in relation to genome size were recently summarized [11]. The traditional view suggests that more than 90% of human genome are nonfunctional and therefore regarded as "junk DNA", whereas ENCODE project recently argued that up to 80% of genome sequences have functional roles [2,12]. Of course, the two opinions are also being on the road for heated debate. Here, we analyzed the genome sequences for 32 representative eukaryote species and roughly illustrated their comparisons on genome size, GC content, and relative proportions of intergenic regions, exons and introns (Fig. 1). Unsurprisingly, an intuitional correlation between genome size and fraction of intergenic regions could be drawn out. Additionally, the proportions of exons and introns show consistent changes more or less.

2. Well-characterized features within genome sequence

Although it is impossible to be completely verified, the conserved features of DNA sequence would exist for corresponding to various biological functions, while some of them are already known but some unknown yet. On the basis of this supposition, we are able to perform *in silico* analysis of DNA sequences for functional investigations. On the whole, features of DNA sequence in eukaryotic genomes could be routinely categorized into two classes, including the compositional properties and functional signals (Fig. 2).

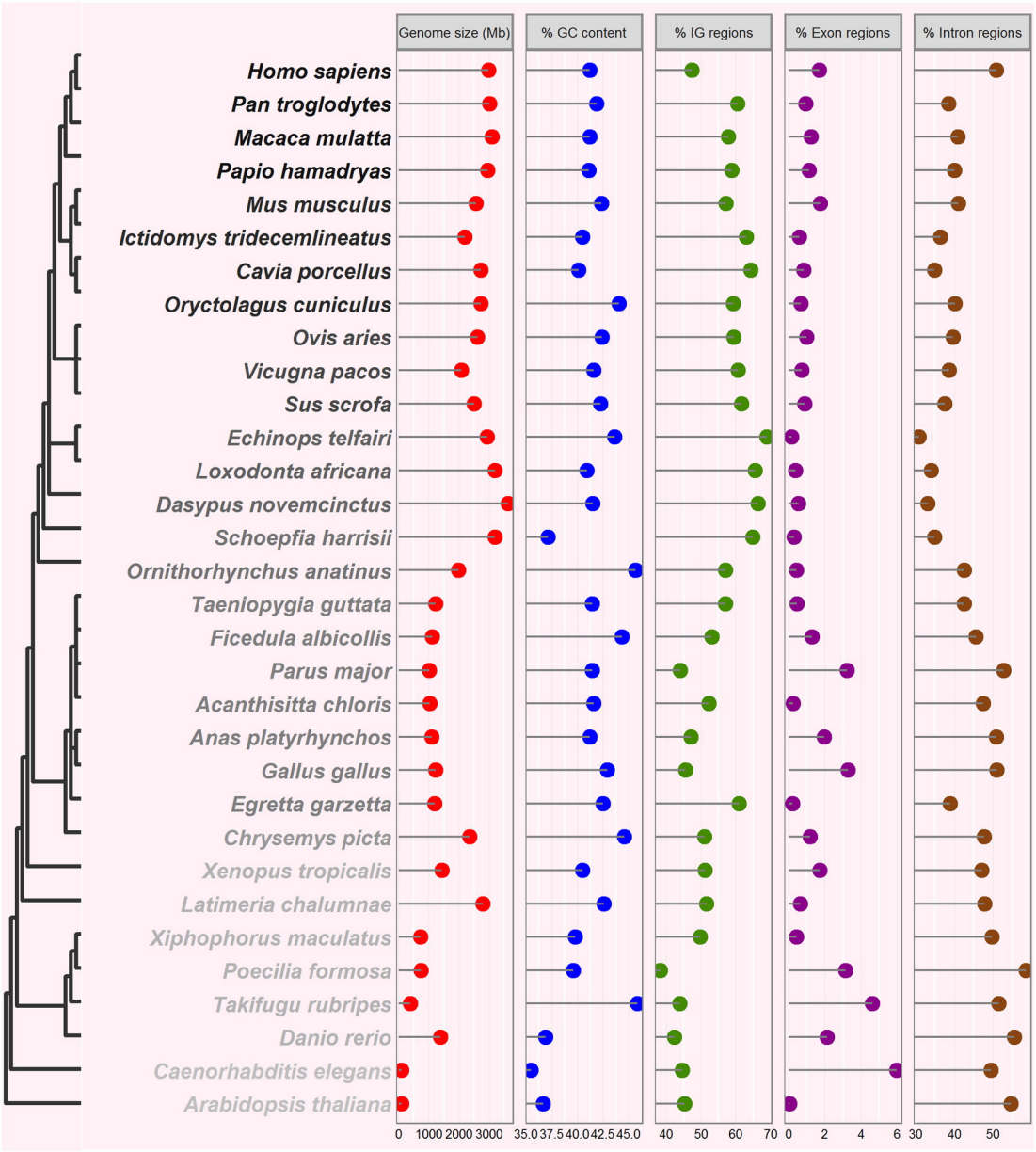

Fig. 1. Architecture of eukaryotic genomes. A total of 32 representative species are included for comparatively illustrating the genome size, GC content, as well as respective proportions of intergenic regions (IG), exons and introns. In brief, all five indices were generated by the dissection of annotation information of reference genome (in GFF format) downloaded from NCBI (March, 2016); and these steps were performed using in-house scripts written in Python language. Additionally, the screenshot of NCBI taxonomic tree is employed to show the phylogenetic relationships among species, in which the full Latin scientific names of species were used.