Computational and Structural Biotechnology Journal 14 (2016) 200–206

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Computational methods for predicting genomic islands in microbial genomes

Bingxin Lu *, Hon Wai Leong *

Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417, Republic of Singapore

ARTICLE INFO

Article history:
Received 31 January 2016
Received in revised form 1 May 2016
Accepted 3 May 2016
Available online 07 May 2016

Keywords:
Pathogenicity islands
Sequence composition
Genome segmentation
Comparative genomics
Outlier detection

ABSTRACT

Clusters of genes acquired by lateral gene transfer in microbial genomes, are broadly referred to as genomic islands (GIs). GIs often carry genes important for genome evolution and adaptation to niches, such as genes involved in pathogenesis and antibiotic resistance. Therefore, GI prediction has gradually become an important part of microbial genome analysis. Despite inherent difficulties in identifying GIs, many computational methods have been developed and show good performance. In this mini-review, we first summarize the general challenges in predicting GIs. Then we group existing GI detection methods by their input, briefly describe representative methods in each group, and discuss their advantages as well as limitations. Finally, we look into the potential improvements for better GI prediction.

1. Introduction

Lateral gene transfer (LGT) is the transfer of genes from one organism to another in a way that is different from reproduction. Its ability to facilitate microbial evolution has been recognized for a long time [1]. Despite the ongoing debate about its prevalence and impact [2–4], the accumulation of evidence has made LGT widely accepted as an important evolution mechanism of life, especially in prokaryotes [5,6]. As a result of LGT, recipient genomes often show mosaic composition, in which different regions may have originated from different donors. Moreover, some DNA sequences acquired via LGT appear in clusters. These clusters of sequences were initially referred to as *pathogenicity islands* (PAIs) [7], which are large virulence-related inserts present in pathogenic bacterial strains but absent from other non-pathogenic strains. Later, the discoveries of regions similar to PAIs but encoding different functions in non-pathogenic organisms lead to the designation of *genomic islands* (GIs) [8]. GIs are then found to be common in both pathogenic and environmental microbes [9].

Specifically, a GI is a large continuous genomic region arisen by LGT, which can contain tens to hundreds of genes. The size of known GIs varies from less than 4.5 kb to 600 kb [3]. Laterally acquired genomic regions shorter than a threshold are also called *genomic islets* [10,11]. GIs often have phylogenetically sporadic distribution. Namely, they are present in some particular organisms but absent in several closely related organisms. As shown in Fig. 1, GIs have several other well-known features to distinguish them from the other genomic regions [10,12,13], such as different sequence composition from the core genome, the presence of mobility-related genes, flanking direct repeats (DRs), and specific integration sites. For example, tDNA (tRNA or tmRNA gene) is well known as a hotspot for GI insertion [11,14]. However, not all these features are present in a GI, and some GIs lack many of these features. As a consequence, GIs were also considered as a superfamily of mobile elements with core and variable structural features [15].

In addition to the restricted GI definition in [16], GIs are often seen as a broad category of mobile genetic elements (MGEs) [17]. They can be further grouped into subcategories by mobility: some GIs are mobile and hence can further transfer to a new host, such as integrative and conjugative elements (ICEs), conjugative transposons and prophages; but other GIs are not mobile any more [10]. GIs can also be classified by the function of genes within as follows: PAIs with genes encoding virulence factors; *resistance islands* (REIs) with genes responsible for antibiotic resistance; *metabolic islands* with genes related to metabolism; and so on [9]. However, the latter classification may not be definite since the functions of genes within GIs may not be clear-cut in practice.

GIs play crucial roles in microbial genome evolution and adaptation of microbes to environments. As part of a flexible gene pool [18], the acquisition of GIs can facilitate evolution in quantum leaps, allowing bacteria to gain large numbers of genes related to complex adaptive functions in a single step and thereby confer evolutionary advantages [9,10]. Remarkably, the genes inside GIs can influence a wide range of important traits: virulence, antibiotic resistance, symbiosis, fitness, metabolism, and so on [9,10]. In particular, PAIs can carry many genes contributing to pathogen virulence [12,13,19], and potential vaccine candidates were suggested to locate within PAIs [20]. Thus, the accurate

* Corresponding authors. Tel.: +65 6516 2903; fax: +65 6779 4580.
E-mail addresses: bingxin@comp.nus.edu.sg (B. Lu), leonghw@comp.nus.edu.sg (H.W. Leong).

http://dx.doi.org/10.1016/j.csbj.2016.05.001

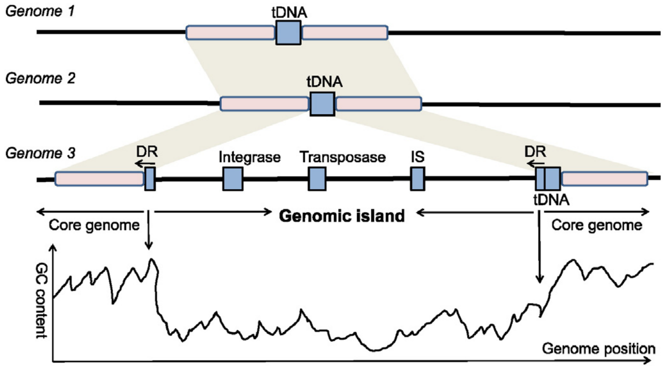

Fig. 1. The schematic representation of several GI-associated features. A GI is often absent in closely related genomes. It may also have atypical compositional characteristics compared with the core genome, such as lower GC content. The presence of several sequence elements is indicative of a GI: flanking conserved regions, DRs, insertion sequence (IS) elements and mobility-related genes encoding integrase and transposase.

identification of GIs is important not only for evolutionary study but also for medical research.

GIs can be predicted by either experimental or computational methods. Herein, we focus on the in silico prediction of GIs: given the genome sequence of a query organism, identify the positions of GIs along the query genome via computer programs alone. Additional input information may also be incorporated, such as the genomes of other related organisms, and genome annotations.

Langille et al. [17] gave a comprehensive review of GI-related features and different computational approaches for detecting GIs. Recently, in 2014, Che et al. [21] presented a similar review for detecting PAIs. Here, we want to provide an up-to-date review of representative GI prediction methods in an integrative manner. Firstly, we highlight the general challenges in predicting GIs. Then, we subdivide existing methods based on input information, and describe their basic ideas as well as pros and cons. We also propose the promising directions for developing better GI detection methods.

2. Challenges in GI prediction

It is a non-trivial task to find laterally transferred regions of relatively small size in a long genome sequence. Two prominent challenges in GI prediction are the extreme variation of GIs and the lack of benchmark GI datasets.

2.1. The extreme variation of GIs

It seems easy to predict GIs given the various well-characterized features associated with it. However, the mosaic nature and extreme variety of GIs increase the complexity of GI prediction [3]. The elements within a GI may have been acquired by several LGT events (probably from different origins) and are likely to have undergone subsequent evolutions, such as gene loss and genomic rearrangement [9]. Consequently, the composition, function and structure of GIs can show various patterns. This can be illustrated by GIs in the same species [22], GIs in Gram-negative bacteria [12], and GIs in both Gram-positive and Gram-negative bacteria [12,15]. The diversity of GIs prevents an effective way of integrating multiple features for prediction. Choosing only a few features as predictors may discard lots of GIs without those features. Even if the fundamental property of GIs, the lateral origin, can be used for prediction, it is still challenging since LGT itself is difficult to ascertain [23].

2.2. The lack of benchmark GI datasets

There are still no reliable benchmark GI datasets for validating prediction methods or supervised prediction. With more GIs being predicted and verified, several GI-related databases have been deployed and regularly updated, such as Islander [24], PAIDB [25], and ICEberg [26] (Table 1). However, these databases are mainly for *specific kinds* of GIs, such as tDNA-borne GIs (GIs inserted at tRNA or tmRNA gene sites), PAIs, and ICEs. There are also two constructed GI datasets based on whole-genome comparison [15,27] (Table 1), which were used as training datasets for machine learning methods. But the scale of these datasets is still not large enough, and their reliability has not been verified by convincing biological evidence.

3. GI prediction methods

In spite of the above challenges, previous methods have made considerable progress in GI prediction. They usually use two most indicative features of the horizontal origin of GIs: biased sequence composition and sporadic phylogenetic distribution. Based on the two features, these methods roughly fall into two categories: *composition-based methods* and *comparative genomics-based methods* [17].

For ease of discussion, we categorize GI prediction methods into two large groups based on the number of input genomes: *methods based on one genome* and *methods based on multiple genomes*. Methods in the former group are often composition based, while methods in the latter group are usually comparative genomics-based. We also include *ensemble methods* which combine different kinds of methods and *methods for incomplete genomes* which predict GIs in draft genomes (in the form of contigs or scaffolds rather than complete whole genome sequence). Fig. 2 shows an overview of the methods included in this paper. For reference, we list available programs which are discussed under each category in Table 2.

3.1. Methods based on one genome

Most methods based on one genome utilize sequence composition to identify GIs, but several methods based on GI structural characteristics have also been developed. According to the units for measuring genome composition, composition-based methods can be divided into *methods at the gene level* and *methods at the DNA level*. In the following sub-sections, we present the basic idea of composition-based methods before discussing methods at the gene and DNA level separately.

The major assumption of composition-based methods is that mutational pressures and selection forces acting on the microbial genomes may result in species-specific nucleotide composition [52]. Thus, a laterally transferred region may show *atypical composition* which is distinguishable from the average of the recipient genome. Under this assumption, most compositional methods try to choose certain sequence characteristics as discrimination criteria to measure the compositional differences. Several features have been shown to be good criteria, including *GC content*, *codon usage*, *amino acid usage*, and *oligonucleotide* (*k-mer*) *frequencies* [53]. Based on these criteria, single-threshold methods are often adopted for GI prediction. The atypicality of each gene or genomic region is measured by a score derived from the comparison with the average of the whole genome via similarity measures. The genes or genomic regions with scores below or above a certain threshold (either predefined or dynamically computed) are supposed to be atypical. The consecutive atypical genes or genomic regions are finally merged to get candidate GIs.

3.1.1. Methods based on gene sequence composition

Methods based on gene sequence composition are often designed to detect LGT, or laterally transferred genes [54], and only a few methods are specifically developed to detect GIs. The methods for LGT detection can be utilized to identify GIs by combing clusters of laterally transferred genes, but they are supposed to be less sensitive, since some genes inside a GI may not show atypicality to allow the whole GI being captured. Here we mainly discuss specific methods for GI detection.