Chapter 3.16
Identification of Genomic Islands by Pattern Discovery

Nita Parekh
International Institute of Information Technology Hyderabad, India

ABSTRACT

Pattern discovery is at the heart of bioinformatics, and algorithms from computer science have been widely used for identifying biological patterns. The assumption behind pattern discovery approaches is that a pattern that occurs often enough in biological sequences/structures or is conserved across organisms is expected to play a role in defining the respective sequence's or structure's functional behavior and/or evolutionary relationships. The pattern recognition problem addressed here is at the genomic level and involves identifying horizontally transferred regions, called genomic islands. A horizontally transferred event is defined as the movement of genetic material between phylogenetically unrelated organisms by mechanisms other than parent to progeny inheritance. Increasing evidence suggests the importance of horizontal transfer events in the evolution of bacteria, influencing traits such as antibiotic resistance, symbiosis and fitness, virulence, and adaptation in general. In the genomic era, with the availability of large number of bacterial genomes, the identification of genomic islands also form the first step in the annotation of the newly sequenced genomes and in identifying the differences between virulent and non-virulent strains of a species. Considerable effort is being made in their identification and analysis and in this chapter a brief summary of various approaches used in the identification and validation of horizontally acquired regions is discussed.

INTRODUCTION

Numerous biological events are responsible for the gradual change in the genetic information of an organism over the course of time such as gene conversions, rearrangements (e.g., inversion or translocation), large-scale deletions and insertions of foreign DNA (e.g., plasmid integration, transposition) apart from point mutations. Horizontal Gene Transfer (HGT) is a major event responsible to cause significant alterations in the genome

DOI: 10.4018/978-1-61350-456-7.ch3.16

composition. It is defined as the transfer of DNA between diverse organisms by mechanisms other than direct descent (vertical inheritance). The clusters of genes acquired as a single unit by horizontal transfer are called ''Genomic Islands (GIs)'' and are typically 10 - 200 Kb in size. These horizontally acquired regions are responsible in causing significant alterations in the genome composition and may provide the organism to carry out new functions resulting in adaptation to a changing environment. Any biological advantage provided to the recipient organism by the transferred DNA creates selective pressure for its retention in the host genome and several pathways of horizontal transfer have been established influencing traits such as antibiotic resistance, symbiosis and fitness, virulence and adaptation in general (Koonin *et al*, 2001; Lawrence & Ochman, 2002; Andersson, 2005; Gogarten & Townsend, 2005). For example, HGT has been demonstrated in many pathogenic strains of bacteria and shown to be responsible for its virulence. Thus, depending on their acquired functions these genomic islands are further classified as pathogenicity islands, metabolic islands, secretion islands, resistance islands and symbiosis islands (Lio & Vannucci, 2000).

General Characteristics of Genomic Islands (GIs)

The genomic islands are found to contain some characteristic features shown in Figure 1 which have been exploited for their identification (Dobrindt *et al*, 2004; Juhas *et al*, 2009). They typically contain in their vicinity intact (or residual) mobile genetic elements, such as genes coding for integrases (Int) or transposases that are required for chromosomal integration and excision, are generally found to be flanked by direct repeats (DR) and are sometimes inserted in the vicinity of tRNAs and tmRNAs, commonly referred to as tRNA genes. Typically GIs also carry multiple functional and fragmented insertion sequence (IS) elements for carrying out the transposition event (Dobrindt *et al*, 2004). The identification of these elements basically involves searching various databases of these elements, *viz*., RepBase Update, tRNA database, etc. by pattern matching.

Apart from the structural features observed in the vicinity of a genomic island, these regions also exhibit bias in the nucleotide compositions. In any genome, ancestral (vertically transmitted) genes experience a particular set of directional mutation pressures mediated by the specific features of the replication machinery of the cell, such as the balance of the dNTP pools, mutational biases of the DNA polymerases, efficiency of mismatch repair systems and so on (Lawrence, 1999). As a result each genome exhibits its own unique signatures such as distinct variations in the GC content, dinucleotide relative abundance, variations in the usage of *k*-mer words, codon usage and amino acid usage. Thus, 'foreign' genes acquired through lateral transfer retain the characteristics of the donor genome which may significantly differ from that of the host genome (Lio & Vannucci, 2000, Lawrence & Ochman 1998). Thus variations in the occurrences of patterns of

Figure 1. General characteristics of genomic islands (adapted from Dobrindt et al., 2004).