Biochimie (1993) 75, 409–413

Analytical approaches to genomic evolution

D Sankoff

Université de Montréal, Centre de Recherches Mathématiques, CP 6128, Succursale A, Montréal, Québec, Canada H3C 3J7

(Received 16 December 1992; accepted 29 December 1992)

Summary — We model the non-local mechanisms of genomic evolution and propose methods for studying the evolutionary divergence of species based on these models. Mechanisms include the movement of segments of genomes within a single chromosome (transpositions), the reciprocal translocation of segments between two chromosomes, and the inversion of segments. Each of these is studied in the context of a different type of genomic data. We introduce the theory of phylogenetic invariants for evolutionary inference based on very long macromolecular sequences.

molecular evolution / chromosome inversion / genome shuffling / phylogenetic invariants

Introduction

Genes evolve through the local processes of nucleotide substitution, insertion and deletion. Genomes, which contain all the genetic information of the organism, necessarily evolve when their component genes evolve in this way. In addition, several non-local evolutionary mechanisms also operate at the genomic level. Entire genes, or segments of chromosomes made up of a series of genes, are inserted or removed as a single evolutionary event. Other segments migrate, are 'transposed' from one region of the genome to another. A segment of a chromosome can be inverted: '*xabcdy*' becomes '*xdcbay*', and the '*dcba*' is relocated on the complementary strand of the DNA, so that reading always proceeds in the right direction. In multi-chromosomal organisms, reciprocal translocation can exchange segments between two chromosomes. Genomic comparison, as an approach to inferring evolutionary divergence, must take account of all these processes. In the following sections of this paper, we summarize a number of studies undertaken from this viewpoint.

The study of genomic evolution involves the analysis of very long sequences of DNA (5×10^4–10^9 terms). In contrast to taxonomic studies based on morphological characters or even on the DNA sequences of individual genes (typically 10^2–10^3 terms), the sequences of genomes justify, by their sheer length, stochastic models where the different terms behave like independent and identically distributed random variables. From this derives an interest for new methods for inferring the structure of trees based on parametric models, in preference to non-probabilistic methods like parsimony or hierarchical classification. For the last 5 years, the method of phylogenetic invariants has drawn the attention of biologists and mathematicians. We explore this subject in the last section of this paper.

Mechanisms of genomic evolution

Mathematical models of evolution at the genomic level [1], and the inferential apparatus associated with them, are qualitatively different from the traditional theory of macromolecular sequence comparison [2]. Even if the well-known processes of insertion and especially of deletion of nucleotides have their counterparts at the genomic level, this is not the case for the predominant process, that of the substitution of one nucleotide for another. At the genomic level, other processes take on importance. These mechanisms can involve two or more remote regions of the genome, in contrast to processes like insertion, deletion and substitution of nucleotides, which are all local operations. We will discuss, in a common framework, analyses of the movement of segments of genomes within a single chromosome (transpositions), of the reciprocal translocation of segments between two chromosomes (*eg* [3]), and of the inversion of segments (*eg* [4, 5]).

At the outset, we may ask whether the frequency and the regularity of these processes justify their use as a statistical basis for the evaluation of the similarity, the distance, or the evolutionary divergence between species, in analogy with nucleic acid sequence