Reconstruction of chromosomal structures

Under a certain approximation, the following simplified format can be adopted. *DNA* is considered as a very long sequence in the {*A*,*C*,*T*,*G*} alphabet. Certain regions of the sequence with specific positions have a particular role and are called *genes*. The partly ordered set of genes of DNAs can be considered as *genome*. The region between neighboring genes (*intergenic region*) can also be functional by triggering the activity of the subsequent gene and are thought to contain a *regulatory signal*, which functions as a switch. Regulatory signals can also be included into the genome.

Evolution is any process developing in time. Evolution of genes, regulatory signals, and genomes are particular cases of evolution. For the sake of simplicity, only one property of the evolving object is considered and all other properties are ignored. Here, we ignore the textual content of the gene; in such a case, DNA is represented by a set of conventional segments (e.g., of unit length), each segment represents a gene, while the distance between genes and the content of intergenic regions are also ignored. It is convenient to consider that these regions have zero length, i.e., the endpoints of neighboring genes are merged together. Every gene can be *read* (and their reading is among the fundamental processes of Life) in two directions: from left to right and from right to left. Accordingly, each segment is directed, and hence is a vector. In this context, genome is a graph composed of directed paths and circles. Analysis of such graphs is of mathematical interest. Such graph referred to as a *CC-graph* reproduces the *chromosomal structure* of genome.

Here, evolution consists in relocation of genes; the corresponding events are referred to as *evolutionary*. It is possible to compile a short list of evolutionary events. They can be divided into types with specific costs that are positive numbers. One genome can be transformed into another by a sequence of evolutionary events. In terms of CC-graphs, one CC-graph can be transformed into another using a definite sequence of operations on CC-graphs, each of which corresponds to a natural event.

The first problem: given two CC-graphs, find a sequence of evolutionary events (operations) to transform the first graph into the second with the minimum *total cost*. This minimum cost will be referred to as a *distance* between these CC-graphs. It is important to find not only the distance but also the *minimum* (or *shortest*) *sequence of evolutionary events*. Thus, the solution of the first problem describes a possible evolution of one genome into another at the level of their chromosome structures.

The second problem: given a set of CC-graphs; let them correspond to leaves of a certain tree *S*, which is yet unknown to us. Imagine that all internal nodes (i.e., all nodes except leaves) of the tree *S* are arbitrarily assigned to any CC-graphs. The tree *S* with such an assignment will be referred to as an *arrangement*. Each arrangement is assigned a *cost*, a sum of distances between CC-graphs at the ends of an edge for all edges of the tree *S*. Find a tree *S* and an arrangement of it with the minimum cost. Thus, the solution of the second problem describes a possible evolution of a set of genomes (e.g., in a given species) at the level of their chromosome structures. The solution

arrangement is referred to as the *reconstruction* of chromosome structures defined in the leaves along the whole tree. Both problems are of great interest as well as the extension of their formulations. The tree *S* is called *phylogenetic*. The above sum for all edges of the tree *S* (or a similar value) is called the *functional*.

One of the main aims of phylogenomics is the reconstruction of objects defined in the leaves along the whole phylogenetic tree to minimize the specified cost (functional), which may also include the phylogenetic tree generation itself. Such objects can include nucleotide and amino acid sequences, chromosomal structures, etc. The structures can have any set of linear and circular chromosomes, variable gene composition and include any number of paralogs, as well as any weights of individual evolutionary operations to transform a chromosome structure. Many heuristic algorithms were proposed for this purpose, but there are just a few *exact algorithms with low* polynomial computational complexity among them. The algorithms naturally start from the calculation of both the distance between two structures and the shortest sequence of operations transforming one structure into another. Such calculation per se is an NP-hard problem.

A general model of chromosomal structure rearrangements is considered. Exact algorithms with almost linear or cubic polynomial complexities have been developed to solve the problems for the case of any chromosomal structure but with certain limitations on operation weights. The computer programs are tested on biological data for the problem of mitochondrial or plastid chromosomal structure reconstruction. To our knowledge, no computer programs are available for this model.

Exactness of the proposed algorithms and such low polynomial complexities were proved. The reconstructed evolutionary trees of mitochondrial and plastid chromosomal structures as well as the ancestral states of the structures appear to be reasonable.

Vassily Lyubetsky
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute),
Lomonosov Moscow State University

Publication

Algorithms for reconstruction of chromosomal structures.
Lyubetsky V, Gershgorin R, Seliverstov A, Gorbunov K
BMC Bioinformatics. 2016 Jan 19