Evolutionary Grammars:
A Grammatical Model for Genome Evolution

Jürgen DASSOW† and Victor MITRANA‡

†Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, P.O.Box 4120, D-39016, Magdeburg, Germany.
‡Faculty of Mathematics, Department of Computer Science University of Bucharest, Str. Academiei 14, R-70109 Bucharest, Romania.

Abstract. We introduce a grammatical model for the evolution of genomes on the basis of gene mutations and chromosome mutations and present some properties of such grammars. Few problems which might be biologically relevant are discussed from the computational point of view. On the other hand, the paper suggests a new direction in formal language theory motivated by the common operations of genome evolution.

1 Introduction

The genomes of complex organisms are organized into chromosomes which contain the genes of the organism arranged in linear order. Much of the current data for genomes is in the form of maps which are now becoming available and permits for the first time the study of the evolution of such organisms at the scale of genome.

In the course of its evolution, the genome of an organism mutates by different processes. At the level of individual genes the evolution proceeds by local operations (point mutations) which substitute, insert and delete nucleotides of the DNA sequence. Evolutionary and functional relationships between genes can be captured by taking into considerations only local mutations. However, the analysis of the genomes of some viruses (Epstein-Barr and Herpes simplex viruses, for instance) have revealed that the evolution of these viruses involved a number of large-scale rearrangements in one evolutionary event [4], [9]. These non-local rearrangements include: deletion, inversion, transposition, duplication and crossing-over.

• Deletion cancels a segment of a chromosome.
• Inversion replaces a segment of a chromosome with its reverse DNA sequence.
• Transposition moves a segment to a new location in the genome.
• Duplication copies a segment to a new location.
• Crossing-over results in recombination of genes in a pair of homologous chromosomes by exchanging segments between parental chromatides. This operation can be modelled as a process that exchanges segments at the end of two chromosomes.

Research supported by the Alexander von Humboldt Foundation
It is the premise of this article that DNA, being a language, should be investigated by means of formal language theory. Any biologist may agree that the hierarchical construction of a typical gene may be viewed as a parse tree (see [10] for such a tree and the corresponding grammar). The grammar form is preferable for promoting an abstracted and hierarchical view of the domain. Regular grammars have been written describing very simple genes [1]. Despite one has argued [11] that the genetic language is no more than context-free, these arguments are based on observations restricted just to the amino acid code. Rather recently, the generative grammars have been viewed as models of biological phenomena such as gene regulation [2].

In [10] mutations and rearrangements are modelled grammatically by means of definite clause grammars which can be directly compiled into Prolog code. Our approach is a string grammar with a restricted use of the rules simulating mutations and rearrangements. Thus, these grammars may be interpreted as a model for regulated evolution.

2 The Grammatical Model

We expect the reader to be familiar with some basics of formal language theory. For all undefined notions we refer to [5].

An alphabet is always a finite set. If \( V \) is a set then \( V^* \) is the free monoid generated by \( V \) and \( V^+ = V^* \setminus \{\lambda\} \) is the semigroup generated by \( V \). For an alphabet \( V \) we denote by \( C(V) = \{(w)|w \in V^+\} \), \( (,) \not\in V \).

The \( C \)-length of \( x \in C(V)^+ \) is defined as follows:

\[
l_{gc}(x) = \begin{cases} 
1, & \text{if } x \in C(V) \\
l_{gc}(y) + 1, & \text{if } x = yw, w \in C(V) 
\end{cases}
\]

An evolutionary grammar is a construct

\[
EG = (V, GM, CE, CO, A)
\]

where

- \( V \) is an alphabet (the set of nucleotides).
- \( GM \subseteq \{\text{Sub}, \text{Ins}, \text{Del}\} \) (the set of gene mutations: substitutions, insertions, deletions, respectively)
  - \( \text{Sub} \) is a subset of \( V \times (V \cup \{\}\) \times (V \cup \{\}) \times V \)
  - \( \text{Ins}, \text{Del} \) are subsets of \( V \times (V \cup \{\}) \times (V \cup \{\}) \setminus (V \times \{\} \times \{\}) \)
- \( CE \subseteq \{\text{CDel}, \text{Inv}, \text{Trans}, \text{Dupl}\} \) (the set of chromosomes evolutions: deletions, inversions, transpositions, duplications, respectively)
  - \( \text{CDel} \) and \( \text{Inv} \) are finite subsets of \( C(V)^+ \)
  - \( \text{Trans} \) and \( \text{Dupl} \) are finite subsets of \( (C(V)^+)^3 \)