A Mathematical Model of Evolutionary Systems

Adrian ATANASIU
Faculty of Mathematics, University of Bucharest
Str. Academiei 14, 70109, Bucharest, ROMANIA

Abstract. A new model of evolutionary system (ES for short) is constructed, based on definitions and notation from [2]. Some results are obtained, namely any language can be generated by a darwinist ES, by a 3-ES. Also, a very promising concept, that of halting ES is proposed; we believe that it will have many further applications.

1 Preliminaries

The genomes may be viewed as consisting of chromosomes with genes arranged in linear order. In the course of its evolution, an organism is transformed by different processes. At the level of a species, the evolution proceeds by local operations which insert and/or delete elements of the sequence. The transition from one species to another can be defined using other operations (called in this paper - mutations). Treating these processes by using formal languages raised the possibility to generalize and investigate the structural information contained in biological sequences and - more of that - to see the depending relations in species evolution. This idea was introduced first at all in [3] and used later in many applications of modelling by grammar formalisms. This paper uses the basic definitions and notations from [2], with some modifications; namely,

- the substitution rule is ignored;
- the mutation rule has a new definition;
- the transition between species is possible only by mutations; evolution rules can modify only organisms inside of species;
- the population of a species is defined as a set, not as a word; so, the notion of organism becomes more clear;

The nature preserves only some of the organisms created, which are able to survive when operations of evolution and mutations give them better abilities of life. In order to in-
crease the reliability of capturing this feature of the process of evolution life we take into consideration only those organisms which lie in a given language called support.

2 Basic definitions

We expect the reader to be familiar with the basics of formal language theory. For all undefined notions we refer to [4].
An alphabet is always a finite and nonempty set. We denote by $V^*$ the free monoid generated by the alphabet $V$ under catenation, by $\epsilon$ the empty string and by $V^+$ the free semigroup generated by $V$, i.e. $V^+ = V^* - \{\epsilon\}$.

Let $x$ be an arbitrary string.

- The length of $x$ is denoted by $|x|$.
- The set of all suffixes of $x$, including the empty string, is denoted by $Suf(x)$. Moreover, $Suf(A) = \bigcup_{x \in A} Suf(x)$.

We proceed to the main definition of this paper. Each component of the defined system is accompanied by some informal explanation that attempt to suggest a parallel between the formal model and the natural process of evolution.

**Definition 1** Let $m \geq 1$ be a positive integer. An $m$—evolutionary system ($m$—ES for short) is a construct

$$\Gamma = (V, S_1, S_2, \ldots, S_m, (x_1, x_2, \ldots, x_m))$$

where:

- $V$ is the alphabet of the system;
- $S_1, S_2, \ldots, S_m$ are the species of the system, a species being defined as a triple of the form

$$S_i = (L_i, E_i, M_i), (1 \leq i \leq m)$$

with:

- $L_i$ being pairwise disjoint subsets of $V^*$. They are called the supports of species;
- $E_i = (Ins_i, Del_i)$ being the set of point evolutions rules (insertions and deletions respectively)

$$Ins_i, Del_i \subseteq (V \cup \{\epsilon\}) \times V \times (V \cup \{\epsilon\})$$

- $M_i(i \geq 1)$ being finite sets of mutation rules defined as:

$$M_1 = \emptyset, M_i = \{(Q_1, \ldots, Q_{i-1}) \mid Q_j \subseteq L_j \cup \{\epsilon\}, 1 \leq j \leq i - 1\}$$

- $x_1, \ldots, x_m$ are sets with at most one element, called initial organisms of the system. They can be $\emptyset$ (the species are not created) or $x_i \subseteq L_i \cup \{\epsilon\}$. 

2
If \((x_1, x_2, \ldots, x_m) = (\{\alpha\}, \emptyset, \ldots, \emptyset)\), the system is called darwinist.

A configuration in \(\Gamma\) is a \(m\) - tuple of the form \((P_1, \ldots, P_m)\) where for each \(i\), \(P_i \subseteq V^*\) is called population of the species \(i\). The initial configuration is \((x_1, \ldots, x_m)\).

For two configurations \(w = (U_1, \ldots, U_m)\) and \(z = (V_1, \ldots, V_m)\), we define the relation \(w \Rightarrow z\) given by the next algorithm:

Algorithm \(\text{derivation}(w,z)\):

\begin{algorithm}
\begin{enumerate}
\item for \(i := 1\) to \(m\) do \(V_i := \emptyset\) endfor;
\item for \(i := 1\) to \(m\) do
\begin{enumerate}
\item for every \((a, c, b) \in \text{Ins}_i\) do
\begin{enumerate}
\item for every \(uabv \in U_i\) do
\begin{enumerate}
\item if \(uacbv \in L_i\) then begin
\begin{enumerate}
\item \(V_i := V_i \cup \{uabv, uacbv\}\);
\item \(U_i := U_i \setminus \{uabv\}\);
\end{enumerate}
\end{enumerate}
\item endfor;
\end{enumerate}
\item endfor;
\end{enumerate}
\item for every \((a, c, b) \in \text{Del}_i\) do
\begin{enumerate}
\item for every \(uacbv \in U_i\) do
\begin{enumerate}
\item if \(uabv \in L_i\) then begin
\begin{enumerate}
\item \(V_i := V_i \cup \{uabv, uacbv\}\);
\item \(U_i := U_i \setminus \{uabv\}\);
\end{enumerate}
\end{enumerate}
\item endfor;
\end{enumerate}
\item endfor;
\end{enumerate}
\item \(V_i := V_i \cup U_i\);
\item end;
\item endfor;
\end{enumerate}
\item for \(i := 1\) to \(m\) do
\begin{enumerate}
\item if \(B_i \neq \emptyset\) then
\begin{enumerate}
\item for every \((Q_1, \ldots, Q_{i-1}) \in B_i\) do
\begin{enumerate}
\item repeat as much as possible
\begin{enumerate}
\item if exists \((x_1, \ldots, x_{i-1})\) with \(x_j \in V_j \cap Q_j (1 \leq j \leq i - 1)\) and \(x_1 \ldots x_{i-1} \in L_i \setminus V_i\) then begin
\end{enumerate}
\item endfor;
\end{enumerate}
\item endfor;
\end{enumerate}
\item endfor;
\end{enumerate}
\end{enumerate}
\end{algorithm}
\[ V_i := V_i \cup \{x_1 \ldots x_{i-1}\}; \ V_j := V_j \setminus \{x_j\}, 1 \leq j \leq i - 1 \]

end;
end repeat;
endfor;
endfor;
end derivation.

Comments:

The goals of the second step are:
• Any individual organism of a species participates in at most one modification;
• Any individual organism remains in next generation near the new organism generated by itself using a insertion or deletion rule;
• The order of applying evolution rules for each species are \textit{Ins-Del} (the rules of insertion are used before that of deletion).

The goals of the third step are:
• Any mutation (element of \(B_i\)) can generate an organism from a new species, if that is not already there. If the rule can be applied, the organisms components involved in mutation are erased from populations of their own species. This assures the developing of a new species.
• In one moment, an organism can participate in generation only one organism from a superior species; for other participations, it must be recreated first.
• The passing from a population to another (the species evolution) can be obtained only by mutations; the rules of insertion and deletion assure developing inside of species. This is the main difference from [2].
• In a darwinist system, every new species begins its own developing only when an organism in that species is obtained by a mutation.

The reflexive and transitive closure of relation \(\Rightarrow\) is denoted by \(\Rightarrow^*\).

Observations:
1. If \(M_2 \neq \emptyset\) then \(M_2 = \{\epsilon\}\).
2. How long \(P_k = \emptyset\), no mutation rule from \(M_i, i > k\) can be applied.
3. If for a species \(S_i, Ins_i\) or \(Del_i\) contains more than one rule, they will be applied randomly. If we establish one order, then can exists some rules impossible be be used. One possibility to solve that is to assure if in one moment the order of rules application is \((1, 2, \ldots, n)\), in the next moment the order will be \((2, 3, \ldots, n, 1)\) (a circular permutation of rules).
If there is a species $S_i$ with $(a, x, b) \in Ins_i$, $(a, y, b) \in Del_i$, then the involution rule can’t be applied.

The population of an $m - ES \Gamma$ can be defined in two ways:

1. **Total** population;

   For each $i(1 \leq i \leq m)$ $L(S_i)$ contains all words $\alpha \in L_i$ for which there is a derivation
   $$(x_1, \ldots, x_m) \Rightarrow^* (P_1, \ldots, P_m), \quad \alpha \in P_i.$$ 

   Then, the total population is defined as:
   $$L_t(\Gamma) = L(S_1) \times L(S_2) \times \ldots \times L(S_m)$$

2. **Historical** population;

   $L_h(\Gamma) = \{(P_1, P_2, \ldots, P_m) \mid (x_1, \ldots, x_m) \Rightarrow^* (P_1, \ldots, P_m)\}$

   Obviously, $L_h(\Gamma) \subseteq L_t(\Gamma)$.

**Definition 2** The language $L \subseteq A^*$ is generated by $m - ES \Gamma = (V, S_1, \ldots, S_m, (x_1, \ldots, x_m))$ iff there is $i(1 \leq i \leq m)$ with $L = L(S_i) \cap A^*$.

Let $\Gamma$ be an $m - ES$ system and $I = \{i_1, \ldots, i_p\} \subseteq \{1, \ldots, m\}(i_1 < i_2 < \ldots < i_p)$ a nonempty set of indexes. On defines the $I - projection$ of total/historical population in $\Gamma$ as the language:

$$pr(I(L_t(\Gamma))) = L(S_{i_1}) \times L(S_{i_2}) \times \ldots \times L(S_{i_p})$$

respectively

$$pr(I(L_h(\Gamma))) = \{(P_{i_1}, P_{i_2}, \ldots, P_{i_p}) \mid (x_1, \ldots, x_m) \Rightarrow^* (P_1, \ldots, P_{i_1}, \ldots, P_{i_p}, \ldots, P_m)\}.$$ 

Particularly, for every $i(1 \leq i \leq m)$

$$L(S_i) = pr_{\{i\}}(L_t(\Gamma)) = pr_{\{i\}}(L_h(\Gamma)).$$

**Example 1** Let’s take the evolutionary system $\Gamma = (V, S_1, S_2, S_3, S_4, (\{\epsilon\}, \{\epsilon\}, \{\epsilon\}, \{\epsilon\}))$ where:

$L_1 = a^*, L_2 = b^*, L_3 = c^*, L_4 = \{a^nb^nc^n \mid n \geq 0\}$

$Ins_1 = \{(\epsilon, a, \epsilon)\}, Ins_2 = \{(\epsilon, b, \epsilon)\}, Ins_3 = \{(\epsilon, c, \epsilon)\}, Ins_4 = \emptyset$

$Del_1 = \emptyset, Del_2 = \emptyset, Del_3 = \emptyset, Del_4 = \emptyset$

$M_1 = \emptyset, M_2 = \emptyset, M_3 = \emptyset, M_4 = \{(a^*, b^*, c^*)\}$

A possible derivation in this system can be:

$$((\epsilon), (\epsilon), (\epsilon), (\epsilon)) \Rightarrow ((\epsilon), (\epsilon), (\epsilon), (\epsilon, abc)) \Rightarrow ((\epsilon, a), (\epsilon, b), (\epsilon, c), (\epsilon, abc)) \Rightarrow ($$

$((\epsilon, a^2), (\epsilon, b^2), (\epsilon, c^2), (\epsilon, abc, a^2b^2c^2)) \Rightarrow$

$((\epsilon, a^2b^2c^2), (\epsilon, abc, a^2b^2c^2)) \Rightarrow \ldots$

In this case, $L_t(\Gamma) = L_h(\Gamma) = L_1 \times L_2 \times L_3 \times L_4$.

The language $\{a^nb^nc^n \mid n \geq 0\}$ is generated in this system by species $S_4$. 


The systems $m - ES \Gamma$ and $n - ES \Gamma'$ are $(p,t)$-equivalent ($(p,h)$-equivalent) if there are two subsets of indexes $I \subseteq \{1,2,\ldots,m\}, J \subseteq \{1,2,\ldots,n\}$ with

$$pr_I(L_s(\Gamma)) = pr_J(L_s(\Gamma')), s \in \{t,h\}$$

We can suppose $m \leq n$ without loss generality. If $\Gamma$ and $\Gamma'$ are $(m,t)$-equivalent ($(m,h)$-equivalent) on said that $\Gamma$ models totally (historically) $\Gamma'$ (or, similar, $\Gamma'$ is modelled by $\Gamma$).

Lemma 1 Any $m - ES$ can be totally (historically) modelled by an $(m+1) - ES$.

Proof: Is immediate if we add in the system a new species $S_{m+1}$ with $L_{m+1} = Ins_{m+1} = Del_{m+1} = M_{m+1} = \emptyset$. 

3 Generative properties of Evolutionary Systems

We will show here some properties concerning languages generated by evolutionary systems. The problems can be treated more detailed if on consider some special properties of species; these ideas will be considered later, in other paper.

Theorem 1 If $L$ is a language generated by a $m - ES$, then $L$ can be generated by a darwinist $(m+2) - ES$.

With other words, any $m - ES$ can be total modelled by a darwinist $(m+2) - ES$.

Proof: Let $\Gamma = (V,S_1,\ldots,S_m,(x_1,\ldots,x_m))$ be a $m - ES$ system and $L \subseteq V^*_i(V_i \subseteq V)$ be a language with property that there is $k \leq m$ with $L = S(L_k) \cap V^*$. If $\Gamma$ is a darwinist system, the theorem is proved (this system can be augmented with 2 final species having all components empty).

Otherwise, we shall construct the darwinist system

$$\Gamma' = (V',S'_1,\ldots,S'_{m+2},(\emptyset,\emptyset,\emptyset))$$

as follows:

$V' = V \cup \{\$\,\#,\};$

$L'_j = \{w\$\, w \in Suf(x_i), 1 \leq i \leq m\}, L'_2 = \{\$,\#,\}, L'_i = L_{i-2} \cup \{z\#|z \in x_{i-2}\} \cup \{\epsilon\} (3 \leq i \leq m+2);$

$Ins'_1 = \{(\epsilon,a,\$) | a \in V\}, Ins'_2 = \{(\epsilon,\#,\epsilon), \epsilon\}, Ins'_{i-2} = Ins_{i-2} (3 \leq i \leq m+2);$

$Del'_1 = \{(\epsilon,\$,\epsilon)\}, Del'_2 = \emptyset, Del'_i = Del_{i-2} \cup \{(\epsilon,\#,\epsilon),(\epsilon,\#,\epsilon)\} (3 \leq i \leq m+2);$

$M'_i = \emptyset, M'_j = \{(\epsilon)\}, M'_i = M_{i-2} \cup \{(\epsilon)\}$ if $x_i \neq \emptyset$ or $\{\epsilon\}, M'_i = M_{i-2} \cup \{(\epsilon,\epsilon,\epsilon)\}$ otherwise $(3 \leq i \leq m+2).$

As a remark, the theorem remains true if on defines $S'_i = S_{i-2}$ for all $i > k + 2$ (the generation of language $L$ is free from all species defined after $S_{k+2}$). This proof generates all components in order to justify the assertion that $\Gamma'$ models totally $\Gamma$.

In the first stage the intermediare sequence $(\emptyset,\emptyset,\emptyset) \Rightarrow (\emptyset,\emptyset,\emptyset) \Rightarrow (\emptyset,\#\emptyset,\emptyset,\emptyset) \Rightarrow (\emptyset,\emptyset,\emptyset,\emptyset) \Rightarrow (\emptyset,\emptyset,\emptyset,\emptyset) \Rightarrow (\emptyset,\emptyset,\emptyset,\emptyset) \Rightarrow \ldots$
In the second stage, the first species begins generation of words \( w \). Once a word \( w \) with \( w \in x_i \) is generated, the species \( i \) uses a mutation \( (w, \#, \epsilon, \ldots, \epsilon) \) and the new word \( w \# \) is included into \( L(S'_i) \). Now the "flag" characters \( $, \# \) are erased and then the species \( S'_i \) is ready to be generated using the rules from \( \Gamma \).

In parallel, using a similar process with the first phase, the empty word is rebuild in first \( i-1 \) positions (allowing mutation of a new organism for another species).

Language \( L \) is obviously generated by species \( k+2 \) of system \( \Gamma \). \( \square \)

Remark: If a language \( S \) is generated by the component \( p \) of an \( m-ES \), we can consider, without loss generality, that \( p=m \) (species \( S_{p+1}, \ldots, S_m \) can be ignored because they have no contribution in developing of species \( p \)).

**Theorem 2** Any language \( L \) can be generated by an \( 3-ES \).

**Proof**: Let \( L \subseteq V^* \) be a language (not necessarily classified Chomsky). On build the evolutionary system \( \Gamma = (V \cup \{\#, $\}, S_1, S_2, S_3, (\{\epsilon\}, \{\epsilon\}, \emptyset) \) where:

\[
L_1 = \{\#\}, L_2 = \{w\$ | w \in V^*\}, L_3 = L \cup \#L \cup \#L;
\]

\[
\text{Ins}_1 = \{(\epsilon, \#, \epsilon)\}, \text{Ins}_2 = \{(\epsilon, $, \epsilon)\} \cup \{(\$, a, \epsilon) | a \in V\}, \text{Ins}_3 = \emptyset;
\]

\[
\text{Del}_1 = \emptyset, \text{Del}_2 = \emptyset, \text{Del}_3 = \{(#, \$, \epsilon), (\epsilon, \#, \epsilon)\};
\]

\[
M_1 = \emptyset, M_2 = \emptyset, M_3 = \{(\#, $L)\}.
\]

Let be \( \alpha \in V^* \). If \( \alpha \notin L \), it cannot arrive in \( L(S_3) \), being rejected by support of \( S_3 \). Also, \( \alpha \) is not in \( L(S_1) \) or \( L(S_2) \) because it has no flag characters \( #, \$. \)

If \( \alpha = a_1 \ldots a_n \in L_3 \), on works as follows:

\[
(\{\epsilon\}, \{\epsilon\}, \emptyset) \Rightarrow (\{\epsilon, \#\}, \{\epsilon, \$\}, \emptyset) \Rightarrow (\{\epsilon, \#\}, \{\epsilon, $, a_1\}, \emptyset) \Rightarrow \\
\Rightarrow (\{\epsilon, \#\}, \{\epsilon, $, a_1, \ldots, a_1 \ldots a_{n-1}\}, \emptyset) \Rightarrow (\{\epsilon\}, \{\epsilon, $, a_1, \ldots, a_1 \ldots a_{n-1}\}, \{\#\alpha\}) \Rightarrow (\{\epsilon, \#\}, \{\epsilon, $, \ldots\}, \{\#\alpha, \#\alpha\}) \Rightarrow (\{\epsilon, \#\}, \{\epsilon, $, \ldots\}, \{\#\alpha, \#\alpha, \alpha\}).
\]

\( \square \)

**Theorem 3** Any evolutionary system is total modelled by an evolutionary system in which, for each species \( S_i \), at least one from next assertions is true:

- one of sets \( \text{Ins}_i, \text{Del}_i \) is empty;
- \( \forall(a, x, b) \in \text{Ins}_i \cup \text{Del}_i \) results \( |a| + |b| \leq 1 \).

**Proof**: Is immediate from proof of Theorem 2. \( \square \)

The behaviour of an \( m-ES \) can be modelled using a device like *Petri Nets* location/transition (for details, see [5]).

So, for every species \( S \) on can build a structure

![Diagram](diagram.png)

where \( M \) is the corresponding set of mutations and \( E \) - the set of evolution rules. By firing transition \( M \), a word is realized by inputs catenation; by firing transition \( E \) two words are realized: the old one (input) and the new one - after application of a rule from \( \text{Ins} \) or \( \text{Del} \).

If the set \( M \) and/or \( E \) are empty, the corresponding transitions are erased.
Example 2: For the 4-ES constructed in Example 1, the Petri net is

![Petri net diagram]

and for 3-ES used in the proof of Theorem 2,

![Petri net diagram]

4 Evolutionary Systems with halting

Let $L$ be a language and $\alpha \in L$ be a fixed word; let us consider a $m$-ES $\Gamma$ which generates $L$ and let’s construct a derivation

$$(\{x_1\}, \ldots, \{x_m\}) \Rightarrow^* (P_1, \ldots, P_m)$$

with the property $\alpha \in P_k$ and this is its first apparition in derivation.

- If for any derivation from $(P_1, \ldots, P_m)$ all species $S_i (1 \leq i \leq m)$ continues to be developed separately (without mutations), on say that system $\Gamma$ is with $\alpha$- halting on mutations;

- If from $(P_1, \ldots, P_m)$ only a finite number of moving are possible, on say that system $\Gamma$ is with $\alpha$- halting.

Theorem 4 For any language $L$ and any fixed word $\alpha \in L$ on can construct a 4-ES with $\alpha$- halting on mutations, which generates sublanguages of $L$.

Proof: Let $L \subseteq V^*$ be a language and $\alpha \in L$ be a specified word. On build a 3-ES system for $L$, like in proof of Theorem 2, where a new species $S_4$ is added. The components of
this fourth species are:
\[ L_4 = \{ \beta \# \alpha \mid \beta \in V^* \}, x_4 = \emptyset, Ins_4 = Del_4 = \emptyset, M_4 = \{ (\epsilon, V^*, \# \alpha) \}. \]

In this way, once the word \( \# \alpha \) is generated in \( L(S_3) \) by mutations, it assures immediate possibility to a new mutation for species \( S_4 \). An effect of this mutation will be \( P_1 = \emptyset \) and therefore - using Observation (2), no further mutation is possible.

Moreover, a historical language of a species receives new organisms only by mutations and - because \( S_4 \) is the first species - it has no mutations.

As a remark, the system \( \Gamma \) can develop further on the species \( S_2 \) chaotically, without influence upon \( L(S_3) \). In \( L(S_3) \) a sublanguage of \( L \) is obtained; its content depends directly by the moment of apparition of \( \alpha \) (in fact, of \( \# \alpha \) in \( L(S_2) \)). Because this apparition is randomly obtained, at different behaviours of system \( \Gamma \), in \( L(S_3) \) different sublanguages of \( L \) can be delivered.

Unfortunately, evolutionary systems with \( \alpha \)– halting can be generated only for very special cases of languages (for example, for finite languages). The motivation is that an insertion operation, if has an organism on which can be applied, cannot be stopped for an infinite language.

References


