

## **Taxonomy in Alife. Measures of Similarity for Complex Artificial Organisms**

Maciej Komosinski and Marek Kubiak

Institute of Computing Science,  
Poznan University of Technology  
Piotrowo 3A, 60-965 Poznan, Poland

Maciej.Komosinski@cs.put.poznan.pl

**Abstract.** In this paper a formal approach to construction of a similarity measure for complex creatures is presented. The simulation model is described, and a Framsticks agent is expressed in a formal way. This helps in defining a dissimilarity measure. Two main ideas are discussed with reference to biology, namely genotypic and phenotypic methods. The holistic phenotypic measure is then proposed, where a fast, heuristic algorithm is used. Examples of its application are shown, including mutation and crossing over analysis, and a clustering tree based on distances between pairs of seven artificial individuals.

### **1 Introduction**

Many research works in the field of artificial life concern studies of evolutionary processes, their dynamics and efficiency. Various measures and methods have been developed in order to be able to analyze evolution, complexity, and interaction in the observed systems. Other works try to understand behaviors of artificial creatures, regarding them as subjects of survey rather than black boxes with assigned fitness and performance. There are some works which employ formal views for these purposes [2,5,14].

Artificial life systems, especially those applied to evolutionary robotics and design [4,6,11], are quite complex and it is difficult to understand the behavior of existing agents in detail. The only way is to observe them carefully and use human intelligence to draw conclusions. Usually, the behavior of such agents is non-deterministic, and their control systems are sophisticated, often coupled with morphology and very strongly connected functionally [12].

Thus for the purposes of studying behaviors and populations of individuals, one needs high-level, intelligent support tools [10]. It is not likely that automatic tools will soon be able to produce understandable, non-trivial explanations of sophisticated artificial agents. However, it is possible to devise automatic measures of similarity which will help in observation of regularities, groups of related individuals, etc.

Similarity can be identified in many ways, including aspects of morphology (body), brain, size, function, behavior, performance, fitness, etc. Whatever definition is used, automatic measure of similarity can be useful for

- optimization to introduce artificial niches by modification of fitness values [7,13],
- studies of evolutionary processes and the structure of populations of individuals,
- studies of function/behavior of agents,
- reduction of the number of agents to a small subset of interesting, diverse, unique individuals,
- inferring dendrograms (and hopefully, phylogenetic trees) based on distances between organisms.

For such reasons we developed a measure of dissimilarity on the realistic agents model of *Framsticks*. Although the model is simpler than biological creatures, it is general enough to observe various properties and difficulties. Devising a good measure of similarity on agents of such complexity is not an easy task, and was not studied so far. The results and conclusions of this long-term work are meaningful for the fields of artificial life, optimization, biology, and robotics. In this paper we report the first step towards such measure.

The paper is organized as follows: section 2 describes the model of simulation and evolution, as well as the formal definition of an agent. In section 3 we focus on the dissimilarity measure, discussing our motivations and biological background. In section 4 a few simple experiments are presented. Section 5 summarizes the work and outlines future goals.

## 2 The model

### 2.1 Simulation

Agents in *Framsticks* are built of body and brain. Body is composed of material points (called *parts* in this paper) connected by elastic *joints*. Brain is made from *neurons* and their elements called *neuroitems* (these are receptors, effectors, and neural connections). For more detailed description of the model mechanics and neural network refer to [9,8,1].

It is possible to use the system for simulation of various processes, including local optimization, evolutionary optimization, coevolution of populations, spontaneous evolution, multi-criteria optimization, etc. Such universality is obtained by the use of a scripting language, which controls the overall architecture of the system as well as more specific issues.

*Framsticks* supports a few genetic representations of agents, including direct and developmental [10]. It is possible to easily “plug in” a new encoding and genetic operators. The fact of diverse genotype languages influences the characteristics of evolutionary process in each case, but finally, each encoding has to produce an agent. Thus it is possible to compare phenotypes which are “compatible” because they are build of a set of standardized components (parts, joints, neurons, and neuroitems).

## 2.2 Formal view

Elements of a framstick agent are characterized by many properties. Morphological elements have mass, friction, stiffness, etc., and control elements – weights, sigmoid coefficients, etc. Body and brain are connected. All these aspects make the definition of a similarity measure difficult; to devise a formal measure it is helpful to mathematically describe an agent.

Let  $P$  denote a non-empty set of parts (material points) of a creature's body. We define several functions which describe physical and biological attributes of parts ( $\mathbf{R}$  denotes the set of real numbers):

- *position*:  $P \rightarrow \mathbf{R}^3$
- *orientation*:  $P \rightarrow [0, 2\pi)^3$
- *mass*:  $P \rightarrow \mathbf{R}_{\geq 0}$
- *volume*:  $P \rightarrow \mathbf{R}_{\geq 0}$
- *friction*:  $P \rightarrow \mathbf{R}_{\geq 0}$
- *ingestion*:  $P \rightarrow \mathbf{R}_{\geq 0}$
- *assimilation*:  $P \rightarrow \mathbf{R}_{\geq 0}$

Furthermore, we introduce  $J$  as a non-empty set of joints, and a function:

- *joint\_connects*:  $J \rightarrow P^2$

which represents connections (arcs) between pairs of parts (vertices). The graph corresponding to morphology must be connected, which means that the creature's body cannot have isolated parts. No arcs joining the same pair of vertices may exist. The properties of joints include:

- *stiffness*:  $J \rightarrow \mathbf{R}_{\geq 0}$
- *rot\_stiffness*:  $J \rightarrow \mathbf{R}_{\geq 0}$
- *stamina*:  $J \rightarrow \mathbf{R}_{\geq 0}$

For brain, we define a set of neurons  $N$ , which may be empty. Properties of a neuron can be defined as functions:

- *force*:  $N \rightarrow [0, 1]$
- *inertia*:  $N \rightarrow [0, 1]$
- *sigmoid*:  $N \rightarrow \mathbf{R}$
- *initial\_state*:  $N \rightarrow \mathbf{R}$
- *placement*:  $N \rightarrow P$

The set  $C$  (which may be empty) concerns neuroitems: elements related to connections between neurons and interfacing parts of the body. This set is divided into pair wise distinct subsets, so that  $C = Const \cup RecG \cup RecT \cup RecS \cup RotMuscle \cup BendMuscle \cup Conn$ . These subsets describe constant signal inputs, receptors (equilibrium, touch, smell), muscles with neural control and connections between neurons.

There are also some properties that these items have (some functions are typical to some specific types of neuroitems):

- *parent\_neuron*:  $C \rightarrow N$
- *connected\_to*:  $Conn \rightarrow N$
- *weight*:  $(Const \cup RecG \cup RecT \cup RecS \cup Conn) \rightarrow \mathbf{R}$
- *on\_part\_placement*:  $(RecT \cup RecS) \rightarrow P$
- *on\_joint\_placement*:  $(RecG \cup RotMuscle \cup BendMuscle) \rightarrow J$
- *strength*:  $(RotMuscle \cup BendMuscle) \rightarrow [0, 1]$
- *rot\_range*:  $RotMuscle \rightarrow [0, 1]$

The two top functions describe the structure of a neural network of a creature. The third one defines weights of connections in the network.

Thus a creature can be described by four sets:  $P$ ,  $J$ ,  $N$ , and  $C$ , and relations between their elements. It can be clearly seen that the structure of an agent is quite complicated – it has different aspects: structural (one graph defining body, another one defining neural network which is located on the body’s graph), geometrical (position and orientation in a 3-D space), and also several parameters describing elements of body and brain.

### 3 Similarity measure

#### 3.1 Biological reference

The measures of similarity (or dissimilarity) are widely adopted in practice by biologists for classification and constructing taxonomies of organisms. Such measures are built based on two kinds of information: genotypic (DNA sequences) [15] and phenotypic (construction and behavior of individuals) [16]. These two streams (apart from phenetics/cladistics difference) can be characterized as molecular and morphological approach.

The first emphasizes the role of a DNA sequence as the true indicator of on-going evolutionary processes. The analyzed DNA fragments (or RNA transcripts/protein sequences) are selected based on knowledge of their relative position in the genome and their function (or the lack thereof). The analyzed sequences are often introns – the non-coding regions of genes, which are assumed to accumulate random mutations non-selectively, according to the defined probabilistic model [3]. However, using this approach with artificial systems may raise a few problem issues. With *Framsticks* system, these are:

- organisms’ genomes are very short compared to living organisms, their length varies from a few characters to a few thousands,
- all formats encode organisms on a very high level: the building blocks are entire elements of an ALifeform body structure,
- difficulties arise in the process of selecting common properties, on basis of which the analysis could be performed,
- there are almost no non-coding, coevolving sequences in the artificial genomes, for which the assumption of a probabilistic model could be made.

Translation of the coding format to the two-state binary model (akin to DNA sequences) would not eliminate these problems.

The latter approach (the morphological one), which entails defining the similarity measure using phenetic resemblance, is more promising. In biology, the overall phenetic resemblance is inferred from many different characters' states – which, if possible, should utilize clearly distinguished and independent properties. This serves to minimize the randomness effect in the (dis)similarity measure. Additional operations in this approach cope with the problems of affine transformations (organism size, scale, and rotations of body parts). The end results are recalculated into an overall measure value, most often using a weighted sum model.

The important question is how to choose the properties of organisms for analysis. Answering this requires delving into the ALife model itself. The numerical similarity analysis and clustering of living organisms require such discrimination because biological perspective is inherently incomplete; an all-inclusive model cannot be built mainly because the principles of the entire system are not known. As we know all the details of ALife individuals, such a holistic approach is possible. Thus such approach in construction of a dissimilarity measure is presented in this work.

### 3.2 Preliminary considerations

The problem of similarity estimation in *Framsticks* is closely related to the problem of isomorphism of graphs. Taking only morphology into consideration, the task would be to find the matching between parts of the two agents. Let  $G_1=(P_1, J_1, joint\_connects_1)$  and  $G_2=(P_2, J_2, joint\_connects_2)$  be the graphs representing bodies of two agents. We assume that  $|P_1| \leq |P_2|$  without loss of generality.

In the common situation when  $|P_1| \neq |P_2|$ , it is not possible to find a mutually single-valued function *matching*:  $P_1 \rightarrow P_2$ . However, one can add a number of artificial points to  $P_1$  forming  $P_1'$  set, so that  $|P_1'| = |P_2|$ , and connect these points with artificial joints so that the graph is connected. Thus the sets  $P_1', J_1'$  and the function *joint\_connects* $_1'$  would be obtained. The desired function *matching*:  $P_1' \rightarrow P_2$  should then maximize the covering of edges in both sets.

However, such approach is not feasible. It requires a few steps where it is unknown how to act (adding points-vertices and joints-edges), and the final similarity value will most probably be sensitive to these arbitrary choices. Furthermore, an exact algorithm working on such sophisticated graph representations would have an unacceptably high computational complexity.

### 3.3 Heuristic algorithm

Following these considerations we constructed a heuristic method. This method tries to match the body structure of two individuals based on the degrees of parts as the main piece of information. The problematic step of adding parts and joints to the smaller organism was abandoned, so the matching function has the form *matching*:  $P_1 \rightarrow M$ , where  $M \subseteq P_2$ . In the overall measure, information about the number of

neurons and neuroitems located on point  $p$  was also used. The corresponding functions are:

$$\begin{aligned} neuron\_count(p) &= |\{n \in N: placement(n)=p\}| \\ nitem\_count(p) &= neuron\_count(p) + |\{c \in C: placement(parent\_neuron(c))=p\}| \end{aligned}$$

To match parts in two organisms, the degree of vertices criterion was used:

$$\begin{aligned} dDeg(matching) &= \sum_{p1 \in P1, p2 \in M: p2=matching(p1)} |degree(p1) - degree(p2)| + \\ &+ \sum_{p2 \in M} |degree(p2)| \end{aligned}$$

The second criterion of similarity considers the numbers of neurons on already matched parts:

$$\begin{aligned} dNeu(matching) &= \sum_{p1 \in P1, p2 \in M: p2=matching(p1)} |neuron\_count(p1) - neuron\_count(p2)| \\ &+ \sum_{p2 \in M} |neuron\_count(p2)| \end{aligned}$$

Sets of parts are sorted by *degree*, then by *nitem\_count* and *neuron\_count*. Then the algorithm is matching parts starting from the highest degree groups and trying to find parts that have similar *nitem\_count* and *neuron\_count*. Starting from the highest degree is likely to preserve the most important parts from being unmatched. In case of ambiguity (when points have the same *degree*, *nitem\_count* and *neuron\_count*), the remaining properties of parts are used as discriminating features (except of *position*).

Based on the constructed function *matching*, two parameters *dDeg* and *dNeu* are estimated. Finally, the dissimilarity between two organisms is evaluated using the weighted sum model:

$$dissimilarity(O_1, O_2) = w_{Deg} \cdot dDeg + w_{Neu} \cdot dNeu$$

where  $w_{Deg}$  and  $w_{Neu}$  are weights of these two criteria. The proposed measure has the following properties:

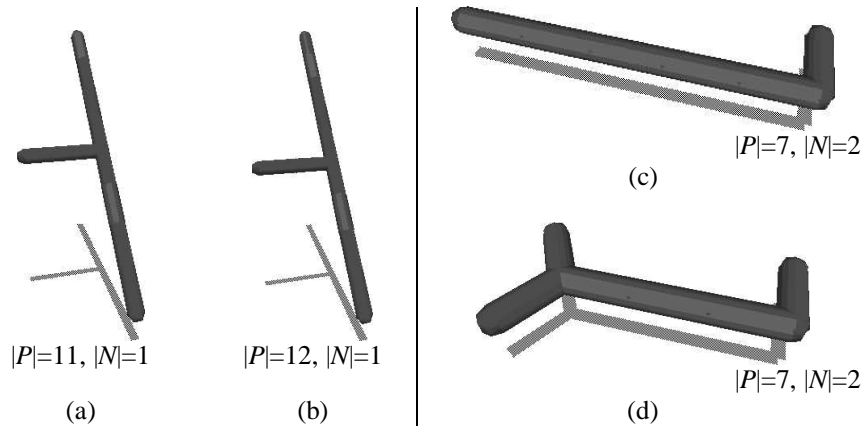
- $dissimilarity(i, i) = 0$ ,
- $dissimilarity(i, j) = dissimilarity(j, i)$ ,
- *dDeg* component holds the triangular inequality.

## 4 Computational experiments

For all experiments, the weights of similarity criteria were  $w_{Deg}=1$  and  $w_{Neu}=0.5$ . These values were adjusted experimentally, constituting an acceptable tradeoff between importance of differences in body and brain.

### 4.1 Mutation

The measure was tested on two random mutations in the “recurrent language” genetic encoding [1,18] (figure 1).

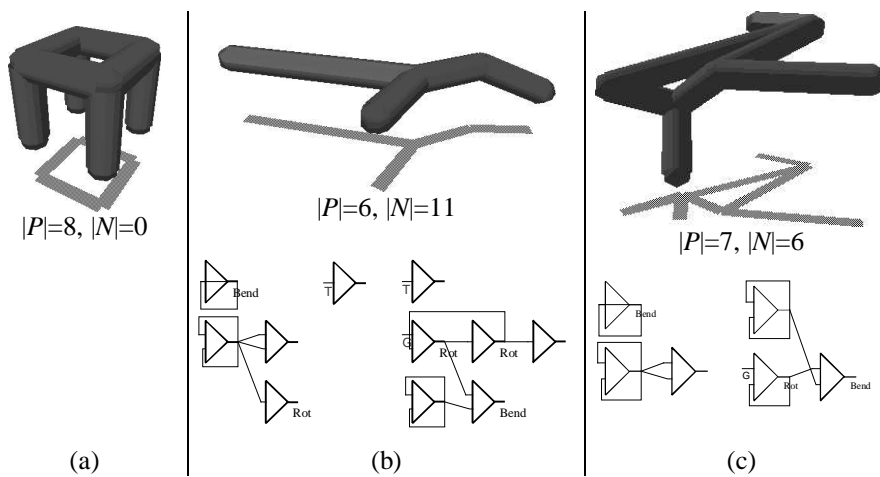


**Fig. 1.** (a) and (c) ancestors, (b) and (d) descendants. Only morphology is shown. In (a) and (b), the branching nodes are in different places.

In both cases, the measure of dissimilarity of ancestors and descendants (which were in genetic terms distant by a single mutation) indicated *dissimilarity* = 2. This small value characterizes well the minor disturbances of the mutation operation. Considering weights used, dissimilarity value of 2 corresponds to the difference of 2 unmatched parts of degree 1, or 4 unmatched neurons.

#### 4.2 Crossover

In the crossover experiment, we combined two agents described by the direct representation, which simply lists all elements of body and brain. The first object was a simple table, which was a human-designed simple object with no neural network. The second one was an individual with a neural network. This individual was the result of a long-term evolution oriented for speed in a land environment (figure 2).



**Fig. 2.** (a) and (b) parents, (c) offspring after crossing over of parents.

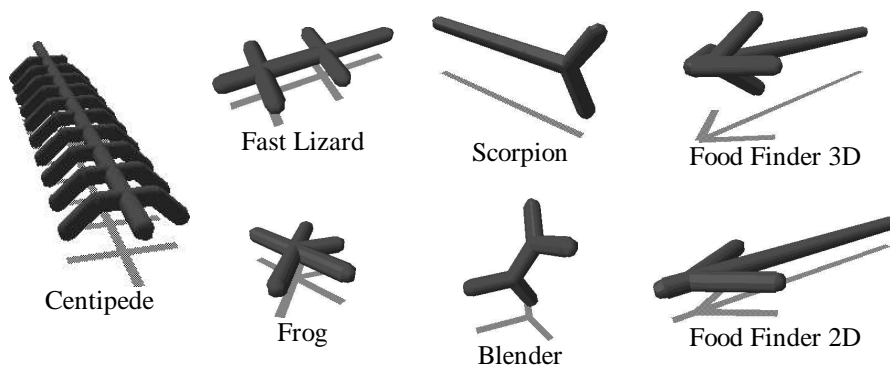
The crossing over operation working on direct representation consists in cutting the parent agents by a 3-D plane and constructing an offspring from two halves of the two parents. It can be seen that the child has some structure parts of parents in its body and brain. The values for the measure are as follows:

$$\begin{aligned} \text{dissimilarity}(a, b) &= 11.5 \\ \text{dissimilarity}(a, c) &= 2 \\ \text{dissimilarity}(b, c) &= 9.5 \end{aligned}$$

It is interesting that  $\text{dissimilarity}(a, b) = \text{dissimilarity}(a, c) + \text{dissimilarity}(b, c)$ , so the child is exactly “between” parents, which reflects the averaging role of the crossing over operator used. This relation holds in the above example, but it is not the general property of the proposed measure. According to the above values, the offspring is very similar to the table and different from the speed-evolved parent. Note that this measure takes into account only structural morphological and neural quantitative properties, but not geometrical aspects (position of parts and length of joints). Thus it can be seen that the child (fig. 2c) has similar structure of the body graph to the parent (fig. 2a).

### 4.3 Clustering

The similarity measure was finally applied to a small example of a clustering problem. Seven diverse creatures were chosen (figure 3).



**Fig. 3.** The seven organisms chosen for the clustering experiment.

The two creatures out of seven can seek food in water. These are food finders; their bodies were designed, and neural networks were evolved. The Centipede and Blender were constructed entirely by a human; Frog’s and Fast Lizard’s morphologies were constructed and brains were evolved; the remaining ones were evolved from scratch.

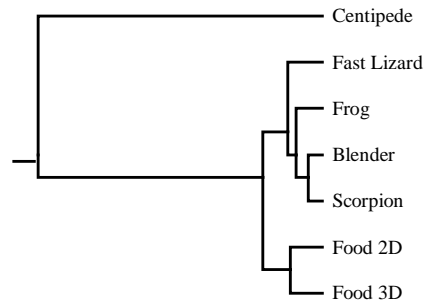
For each pair of these creatures, the dissimilarity was computed (table 1).



|             | Centipede | Fast Lizard | Food 2D | Food 3D | Frog  | Blender | Scorpion |
|-------------|-----------|-------------|---------|---------|-------|---------|----------|
| Centipede   | 0         | 116         | 109.5   | 99.5    | 128.5 | 124.5   | 123      |
| Fast Lizard | 116       | 0           | 11.5    | 19.5    | 13.5  | 12.5    | 10       |
| Food 2D     | 109.5     | 11.5        | 0       | 11      | 20    | 18      | 14.5     |
| Food 3D     | 99.5      | 19.5        | 11      | 0       | 31    | 27      | 25.5     |
| Frog        | 128.5     | 13.5        | 20      | 31      | 0     | 8       | 9.5      |
| Blender     | 124.5     | 12.5        | 18      | 27      | 8     | 0       | 4.5      |
| Scorpion    | 123       | 10          | 14.5    | 25.5    | 9.5   | 4.5     | 0        |

**Table 1.** The symmetric matrix of dissimilarity values of selected creatures.

Based on this matrix, it is possible to construct a clustering tree. UPGMA method was used (unweighted pair group method with arithmetic averages) [16]. It is a simple and intuitively appealing example of the phenetic approach to data summary. UPGMA groups objects that differ least according to the similarity measure without other points of consideration. The obtained hierarchical tree is shown on figure 4. Considering structures of creatures shown on figure 3, it can be seen that the hierarchy is sound.



**Fig. 4.** The resulting clustering tree.

## 5 Conclusions and future work

In this paper, a formal approach to agent analysis was presented. Based on this approach, a structural dissimilarity measure was proposed, and its prototypic applications were shown on some examples.

The measure will be further developed to take into account all the relevant features of an agent, including geometrical and biological properties. Considering geometry creates additional problems, which are also present in biological analyses, namely affine transformations (scaling, rotation, etc.). However, geometry has a great impact on the behavior of agent, functionality, and dynamics of its movement, so it cannot be neglected. Even though, the structural properties considered in this paper are an important holistic property of a creature and cannot be easily modified by single mutations, so it is a good basis for the matching algorithm.

There are strong complexity constraints imposed on such measure, because it is to be used many times. For artificial speciation, the measure has to be computed for all pairs of individuals in each generation. In analysis of structure of populations, it has to be computed once, but the population size is usually much greater than in the case of a typical evolutionary algorithm. Thus only the heuristic approach is possible; exact measures would be too slow to compute for all pairs of individuals. It is

required for the measure to have some good properties, like robustness, simplicity, clarity, etc. Another properties are required if the measure is to be used in a biological, evolutionary context. Our future work will proceed in these directions. Application of this measure to diversify populations of individuals will also be tested.

### Acknowledgements

This work has been supported by the State Committee for Scientific Research, from KBN research grant no. 8T11F 006 19, and by the Foundation for Polish Science, from subsidy no. 11/2001.

### References

1. Adamatzky A., Komosinski M. and Ulatowski S. (2000) Software review: Framsticks, *Kybernetes: The International Journal of Systems & Cybernetics*, **29** (9/10), 1344-1351.
2. Agre P. and Horswill I. (1997) Lifeworld analysis, *Journal of Artificial Intelligence Research*, **6**, 111-145.
3. Avise J.C. (1994) *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.
4. Bentley P. (1999) *Evolutionary design by computers*. Morgan Kaufmann.
5. Bongard J.C. and Paul C. (2000) Investigating morphological symmetry and locomotive efficiency using virtual embodied evolution. In: *Proceedings of the Sixth International Conference on Simulation of Adaptive Behaviour* (ed. by J.-A. Meyer), pp. 420-429. MIT Press.
6. Funes P. and Pollack J.B. (1998) Evolutionary body building: adaptive physical designs for robots, *Artificial Life*, **4** (4, Autumn), 337-357.
7. Goldberg D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Co.
8. Komosinski, M. and Ulatowski, Sz. Framsticks Internet site, <http://www.frams.alife.pl/>
9. Komosinski M. (2000) The World of Framsticks: Simulation, Evolution, Interaction. In: *Virtual Worlds. Lecture Notes in Artificial Intelligence 1834* (ed. by J.-C. Heudin), pp. 214-224. Springer-Verlag.
10. Komosinski M. and Rotaru-Varga A. (2000) From Directed to Open-Ended Evolution in a Complex Simulation Model. In: *Artificial Life VII* (ed. by M.A. Bedau, J.S. McCaskill, N.H. Packard and S. Rasmussen), pp. 293-299. MIT Press.
11. Lipson H. and Pollack J.B. (2000) Automatic design and manufacture of robotic lifeforms, *Nature*, **406** (6799), 974-978.
12. Lund H.H., Hallam J. and Lee W.-P. (1997) Evolving Robot Morphology. In: *Proceedings of IEEE 4th International Conference on Evolutionary Computation*. NJ. IEEE Press. Invited paper.
13. Michalewicz Z. (1996) *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag.
14. Rasmussen S. and Barrett C.L. (1995) Elements of a theory of simulation. In: *Proceedings of the European Conference on Artificial Life (ECAL 95), Lecture Notes in Computer Science*. Springer-Verlag, Berlin.
15. Setubal J. and Meidanis J. (1997) *Introduction to Computational Molecular Biology*. PWS Publishing Company.
16. Sneath P.H. and Sokal R.R. (1973) *Numerical Taxonomy*. Freeman & Co, San Francisco.