# Applications of a similarity measure in the analysis of populations of 3D agents

Maciej Komosinski*

*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland*

**Abstract**

Research in complex collective and multi-agent systems often involves building models of three-dimensional biological life or evolving such structures in virtual environments. Applications stemming from evolutionary design, engineering, robotics, and artificial life require processing of large numbers of such agents that are encoded in some form of a "genotype". However, what is important in evaluation is the "phenotype", i.e. the actual 3D body and its properties. This work introduces a number of ways in which a measure of similarity of 3D agents can support researchers in recognizing the link between the genotype and phenotype spaces, building taxonomies of 3D bodies and automatically selecting representative agents. The measure of similarity employed here is based on phenotypes and places few restrictions on the compared designs, so it can be applied independently of genetic representation.

*Keywords:* evolution, 3D, structure, morphology, similarity

## 1. Introduction

One of the areas where many spectacular examples of complex collective systems exist is biology. In such systems, mutual interactions of multiple agents generate rich and non-linear patterns. When modeling biological life forms and their environments, two important aspects that facilitate complexity are embodiment and embeddedness (situatedness). They highlight the importance of physical interactions between the body and the world, and indicate that the brain is not just an independent entity that processes abstract information. The body and its physics are important factors in a number of fields, especially where motor control, sensory-motor coordination, navigation and mapping, or spatial computing are involved. This concerns practical applications in robotics (evo-devo, morphogenetic, bio-inspired, autonomous), and is also related to body and brain co-evolution, self-assembling and self-replication of structures, and artificial embryogenies and synthetic biology. The body, as a part of an organism, can also be used as an indicator of the complexity of the organism, therefore bodies of creatures can be used as an input for measures of evolved complexity.

The process of evolution, often present in collective systems and usually simplified in computer models as "directed" (with exogenous fitness) or "spontaneous" (with endogenous fitness) evolution, is in fact a search process that generates and processes large numbers of individuals. This work demonstrates how comparing these individuals can provide new information in analyses of evolved or designed virtual bodies. From the formal point of view,

*Corresponding author
*Email address:* maciej.komosinski@cs.put.poznan.pl (Maciej Komosinski)

This paper appeared in *Journal of Computational Science*, 2016.
http://dx.doi.org/10.1016/j.jocs.2016.10.004

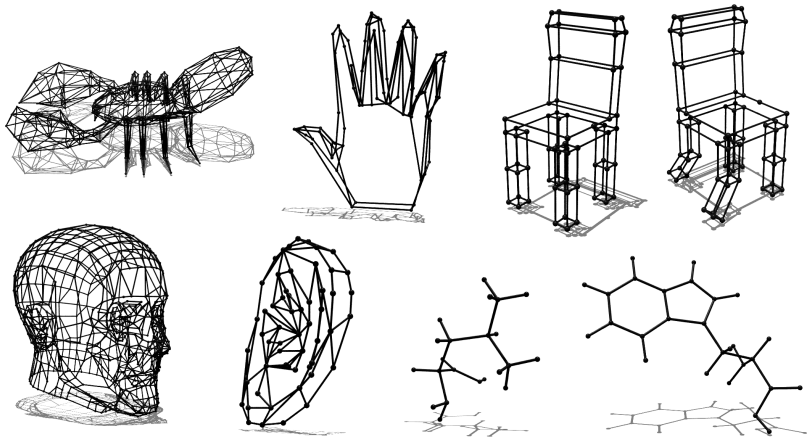

Figure 1: Three-dimensional graph-based bodies compatible with the model considered in this work are used in many disciplines including computer graphics, anatomy, design, engineering, and chemistry.

problems concerning three-dimensional agents are among the most difficult. Since individuals – structures, designs – vary in size, shape, and complexity, their representations consist of variable amounts of information. There exist a large number of genetic representations that are used to encode 3D designs [1, 2, 3], and the level of generality of algorithms and tools employed in this area is low. Since the analysis of 3D agents usually involves human expertise, it is not easy to automate their evaluation and to further support humans in investigation and understanding of the outcomes of optimization.

This article introduces selected applications of a measure which is able to estimate "distance" (i.e., dissimilarity) between a pair of three-dimensional agents. In classical combinatorial problems (consider the similarity of binary strings or permutations as examples), the concept of similarity of solutions has been employed for a long time, because it is relatively easy to come up with a procedure that estimates similarity, and the definition of similarity is usually natural and straightforward. When solutions become complex, there are many possibilities regarding aspects that should be taken into account when the similarity is determined, and the process becomes less obvious. For experiments reported in this work, a heuristic measure has been used [4] to illustrate its usability and the advantages the concept of similarity can bring to the organization of 3D agents.

The particular model of an agent considered in this work can represent any kind of structure where connected vertices are located in a 3D space, as shown in Fig. 1. This includes designs made of a finite number of elements (also active structures and robots [5, 6, 7]), computer graphics objects and meshes, trusses in architecture and structural engineering [1, 8], circulatory and skeletal systems, chemical structures, molecules, and geolocated graphs – e.g. supply, communication, or transmission networks.

Since a similarity measure is going to be applied to optimized agents, it should (1) be able to deal with arbitrary structures, (2) have low computational complexity, and (3) capture aspects of structures that are correlated with fitness in most optimization problems. The third requirement is hard to satisfy *a priori*, but once a measure exists, the relationship between fitness and similarity can be verified on a range of problems. The three requirements are generally not satisfied by dedicated similarity measures which are used for 3D structures in various areas (computer vision [9], bioinformatics and chemical informatics [10]) because those measures are too specific and make additional assumptions regarding structures, or they are too inefficient, or they are based on indirect information about compared structures [11].

To address the three requirements, an original measure of similarity has been proposed [4], and this measure is used in the experiments described here.

This paper is organized as follows: Sect. 2 briefly presents the model of a 3D agent used throughout this work and provides a general idea of the procedure used to estimate dissimilarity. Sect. 3 demonstrates a range of applications of the dissimilarity measure: building taxonomies of agents, visualizing and investigating various aspects of their similarity, reconstructing sequences of random geometric mutations from phenetic dissimilarity matrices, and automatically selecting representative agents based on these distance matrices. Sect. 4 summarizes this work.

## 2. Estimating the dissimilarity of 3D agents

In the experiments presented here, bodies of agents are modeled as undirected graphs. Vertices represent parts of a body, and arcs correspond to joints between these parts. Since the geometry of an agent is essential in most applications, positions of body parts in the 3D space are also considered. In robotics, simulations of life and evolutionary design, agents may be *active* – they may include some kind of a control system. It is assumed that elements of the control system are associated with (or located on) body parts or joints, and since control elements can be connected, they constitute a directed graph.

Additionally, the components of a body (parts, joints) and the control system (sensors, actuators, signal processing units, and their connections) can have a number of properties such as mass, friction and stiffness coefficients, or sensory characteristics. These properties are also part of the model and are taken into account when the dissimilarity of two agents is estimated.

In the model, no assumptions are made regarding the structure of the graphs, their shape, geometry, or occurrence of substructures. Therefore, surface meshes in computer graphics and graphs representing volumes of solids [12] are compatible with this model, as well as skeletal systems and most modularly-designed robots [7]; other implementations of the model can be found in [13, 1] and in [14, chapters 3, 4, 5, and 11].

A detailed description and analyses of the dissimilarity estimation algorithm employed in this work are provided in [4]. The main concept of evaluating the dissimilarity of two models is outlined below. Since estimation of the dissimilarity of graphs is closely related to the problem of isomorphism of graphs, and the exact algorithm that finds the maximal common subgraph is NP-hard, a fast heuristic procedure has been employed.

The heuristic algorithm matches parts of the two structures based primarily on degrees of vertices. In both compared models, vertices are first sorted by their degree, and vertices of the same degree are sorted by the number of associated control units. The matching process starts from groups of parts with the highest degree and the highest number of control units. In order to discriminate between parts, location in space and additional properties of the body and control system are taken into consideration, according to weights adjusted by the researcher.

The steps required to estimate dissimilarity are as follows:

- create models corresponding to the two compared agents,
- match vertices in both models,
- with reference to the above matching, compute individual components of dissimilarity,
- if necessary, aggregate these components into a single value.

The number of vertices in the two models can be different, and in this case, for each extra vertex, a dissimilarity ingredient equal to the dissimilarity of the two maximally dissimilar vertices is added to the appropriate dissimilarity component. As a result of the matching algorithm, four values are obtained: $d_V$ – the difference in the number of vertices, $d_D$ – the difference in degrees of matched vertices, $d_N$ – the difference in control units attached to matched vertices, and $d_G$ – the geometric distance between matched vertices. If required, these four components of dissimilarity can be aggregated into one value, for example using the weighted sum formula:

$$dissimilarity = d_V w_V + d_D w_D + d_N w_N + d_G w_G \tag{1}$$

Using a weighted sum is convenient since the weights are also used in the matching phase. In particular, when $w_G = 0$, the algorithm only estimates the structural difference of the two graphs and geometric aspects of the agents are ignored. When $w_G > 0$, both bodies are aligned in space so that they overlap as much as possible. Details, analyses, and examples of this procedure are provided in [4].

## 3. Applications

The experiments described in this section demonstrate a broad spectrum of applications of the dissimilarity measure. These include clustering agent structures, discovering and understanding evolutionary paths, finding representative agents, and identifying relationships between fitness criteria and body shapes.

The following descriptions focus on the techniques of employing of the measure and do not go into technical details of multi-agent simulations, and do not extensively list all the simulation parameters. All the experiments were performed using Framsticks, a highly configurable and versatile simulation environment [15, 16]. This software is suitable for modeling autonomous agents, complex collective systems, and evolutionary and co-evolutionary processes. It allows for any number of genetic encodings organized in a hierarchy [17, 3], and features two modes of mechanical simulation (accurate rigid body and approximate elastic body) [18, 19]. Framsticks is controlled by a custom scripting language that defines most aspects of the simulation and the user interface. The simulator is available for both desktop and mobile devices and provides a number of user interfaces, from simple graphical ones to advanced command-line and distributed client-server architectures. The open-source SDK [20] includes C++ sources of the dissimilarity estimation algorithm used in the experiments presented here.

### *3.1. Building taxonomies of evolved agents*

Analyzing sets of evolved designs is tedious, difficult for a human and prone to various biases and inaccuracies when performed manually. This was the main motivation for the development of the similarity measure, which could help in automated analyses of large collections of agents. In this study the measure has been applied to process a set of 29 designs shown in Fig. 2. This set contains the best individuals evolved earlier in an experiment where three independent, maximized fitness functions were considered: the height of the center of mass with and without the neural network activated, and the speed of locomotion on a flat terrain [3]. There were 10 independent evolutionary optimization runs for each of the three goals, resulting in a set of the 30 best designs (one highest-fitness individual has been selected from each run). One design, *hp7*, has been excluded from this comparison for better presentation of results. The excluded design has evolved an exceptionally complex body (4353 vertices) while the second most complex design (*h5*) has only 27 vertices and the third one (*hp0*) has 16 vertices. Were this extremely complex design included in the comparison,

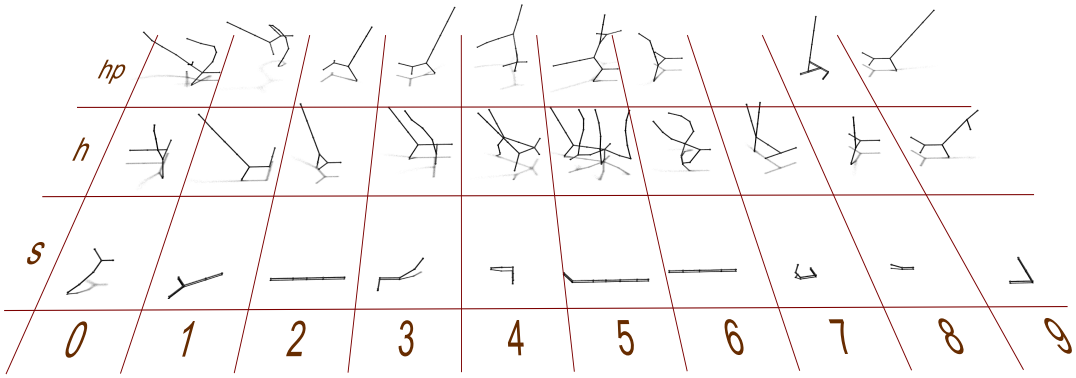


Figure 2: 29 agents evolved in three tasks (maximize passive height, active height and speed).

Table 1: Weights used in the comparative experiment in Sect. 3.1 to compute aggregated dissimilarity.

| Weight | $w_V$ | $w_D$ | $w_N$ | $w_G$ |
|---|---|---|---|---|
| Structure-only value | 5 | 1 | 0 | 0 |
| Geometry-aware value | 5 | 1 | 0 | 1 |

its dissimilarity to all the other simpler structures would be so huge that it would dominate other distances and therefore obscure visualizations of results.

Two distance matrices of size 29 have been computed using the structure-only ($w_G = 0$) and the geometry-aware ($w_G > 0$) dissimilarity estimations. The weights were adjusted to expose the influence of geometrical dissimilarity [4] compared to structure-only dissimilarity [21] as shown in Table 1. Differences in neural networks were not considered at all ($w_N = 0$). The value of $w_V$ corresponds with the highest importance of the difference in complexity (the number of vertices). The intermediate value of $w_D$ reflects the importance of differences in the degree of matched vertices. The value of $w_G$ is 0 or 1, which allows direct observation of the impact of the geometric dissimilarity component.

Both dissimilarity matrices were used to generate hierarchical clustering trees. The unweighted pair group method with arithmetic mean (UPGMA [22]) was applied here. It is a simple agglomerative data clustering technique that uses a sequential algorithm to group objects that differ least (their dissimilarity is lowest).

Fig. 3 shows UPGMA trees for the structure-only and the geometry-aware dissimilarity measures. In both trees, the *h5* body is considered unique because it has many more vertices than other bodies, and the $w_V$ weight is high. The trees reveal three general groups: the bottom group of simple (mostly linear) designs (constituted by agents evolved for speed, "s"), the top group of complex, bush-like structures (*hp0..h6*), and the central group of intermediate designs (*hp2..h7*).

The contents of these three groups is similar in both trees except for the *s5* design. This design was located in the bottom "linear structures" group when only structural properties were evaluated, but when geometry was additionally assessed, *s5* moved to the central "intermediate structural complexity" group. An important advantage of considering the geometrical properties of agents is the increase in resolution of the dissimilarity measure. The structure-only approach was unable to discriminate between a number of bodies (e.g. *hp2*, *hp3*, *hp6*, *hp9*, etc.), while the geometry-aware dissimilarity estimation successfully captured their diversity.

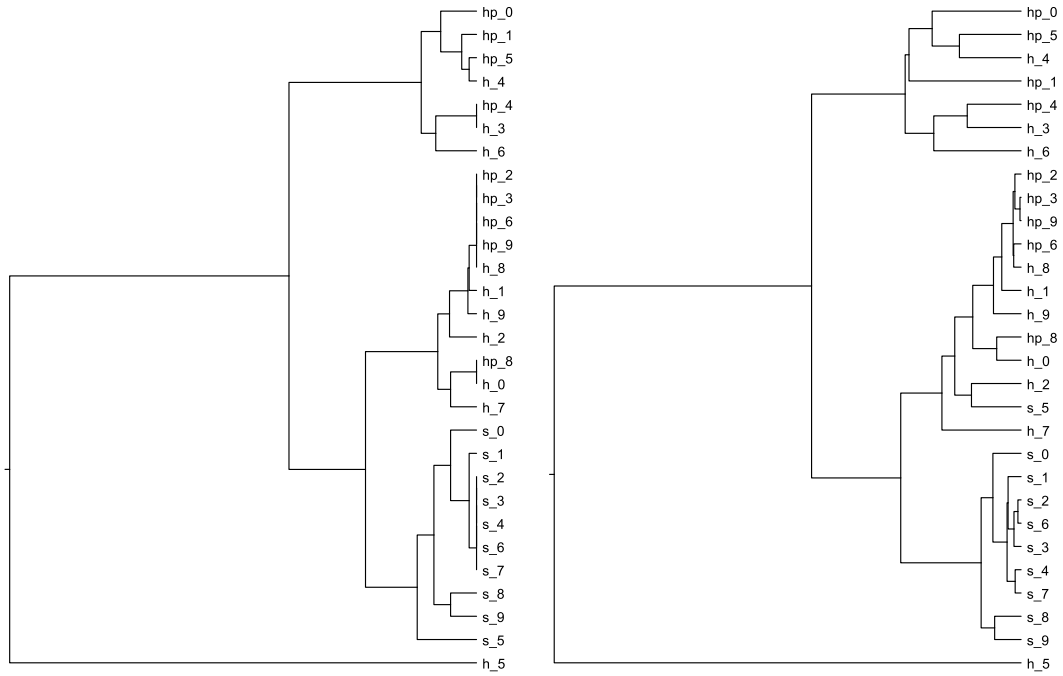


Figure 3: UPGMA trees for the agents shown in Fig. 2 (left: structure-only dissimilarity measure, right: geometry-aware dissimilarity measure).

Another visible difference is the increase in distances in the top and central groups, but not in the bottom group. The top and central groups contain bodies of high and intermediate structural complexity, so geometric differences are highest once these structures are aligned. On the other hand, the bottom group that comprises nearly-linear designs was hardly influenced by the geometrical difference component. This relative increase in the dissimilarity of complex structures is a desired effect, as is the increase in the resolution of dissimilarity where structural aspects are insufficient to discriminate between agents.

Overall, both estimates of dissimilarity (with and without geometrical aspects) proved useful in the analysis of this set of evolved designs, with the geometry-aware approach proving superior in terms of resolution of discrimination.

### *3.2. Visualizing and investigating various aspects of similarity of agents*

The previous subsection focused on the illustration of the geometrical vs. structural components of the dissimilarity measure, and demonstrated that difference on a relatively small, comprehensible example. In this subsection, a larger set of creature models will be processed and more combinations of the weights of the dissimilarity measure components will be employed. This will demonstrate how the adjustment of weights can facilitate the visualization of various aspects of similarity in a set of agents.

An input set consists of creatures evolved to maximize four criteria: vertical position of the center of mass (with and without neural network enabled) and velocity (on land and in water). The vertical position was averaged from measurements taken every 100 simulation steps. For each of these four criteria, 100 creatures have been independently evolved in a steady-state evolutionary algorithm. The most important parameters of the evolutionary

Table 2: Sets of weights of dissimilarity components used for the comparison in Sect. 3.2. Results are shown in Fig. 4.

| | $w_V$ | $w_D$ | $w_N$ | $w_G$ | Fixed vertical axis? |
|---|---|---|---|---|---|
| (a) | 1 | 0 | 0 | 0 | irrelevant |
| (b) | 0 | 1 | 0 | 0 | irrelevant |
| (c) | 0 | 0 | 0 | 1 | no |
| (d) | 0 | 0 | 0 | 1 | yes |
| (e) | 0 | 0 | 10 | 1 | yes |
| (f) | 0 | 0 | 1 | 0 | irrelevant |

process were as follows. The genetic representation was *f1*, also known as *recur* [3, 15, 16]. The gene pool size was 1000. The positive selection mechanism was the tournament of size 3. The evolutionary optimization process consisted of two phases, and each phase ended when one million of non-improving evaluations of new creatures occurred. In the first "exploration" (low selective pressure) phase, the negative selection consisted in removing a random genotype from the gene pool. In the second "exploitation" (high selective pressure) phase, the negative selection was removing the worst genotype from the gene pool.

To evolve more robust creatures, a small amount of non-determinism was included: when a creature was born, the states of neurons in its neural network were initialized randomly. In order to achieve reliable fitness, creatures were evaluated multiple times and performance was averaged. A small penalty was added to the fitness function for each component in body and brain to limit the number of unnecessary components in creatures. The formula to deduct a fraction of the original raw fitness value was adjusted so that the pressure for simplicity was extremely low and did not significantly disturb the fitness landscape. Assuming that "`this`" denotes the currently evaluated creature, the formula was as follows:

```
var fitness=this.velocity; //or this.vertpos for vertical position
var components=this.numparts+this.numjoints+this.numneurons+this.numconnections;
var penalty=1.0-Math.exp(-0.001*components); //exponentially scaled, 0..1
if (fitness>=0)
   return fitness*(1.0-penalty); //positive: penalize, make less positive
else
   return fitness/(1.0-penalty); //negative: penalize, make more negative
```

In Framsticks, body and brain of an agent (the `Model` object) is constructed based on its genotype. Then the model is put into simulation (the `Creature` object) where environmental forces act, so the simulator awaits while the agent stabilizes and then its neural network starts working. Since the morphology of a `Model` does not properly represent evolutionary fitness, in the following analysis morphologies of `Creature` objects are employed from the first moment when they became stable (stationary) after being put in the simulation.

The $4 \times 100$ independently evolved agents (see the video at `https://www.youtube.com/watch?v=lo4vL7gOuYk`) were then mixed into a single unordered set, and six dissimilarity matrices were calculated according to the weights summarized in Table 2. The time to compute a $400 \times 400$ distance matrix on a single 2.4 GHz processor core was approximately 92 seconds when the geometric component was employed, and about 45 seconds in other

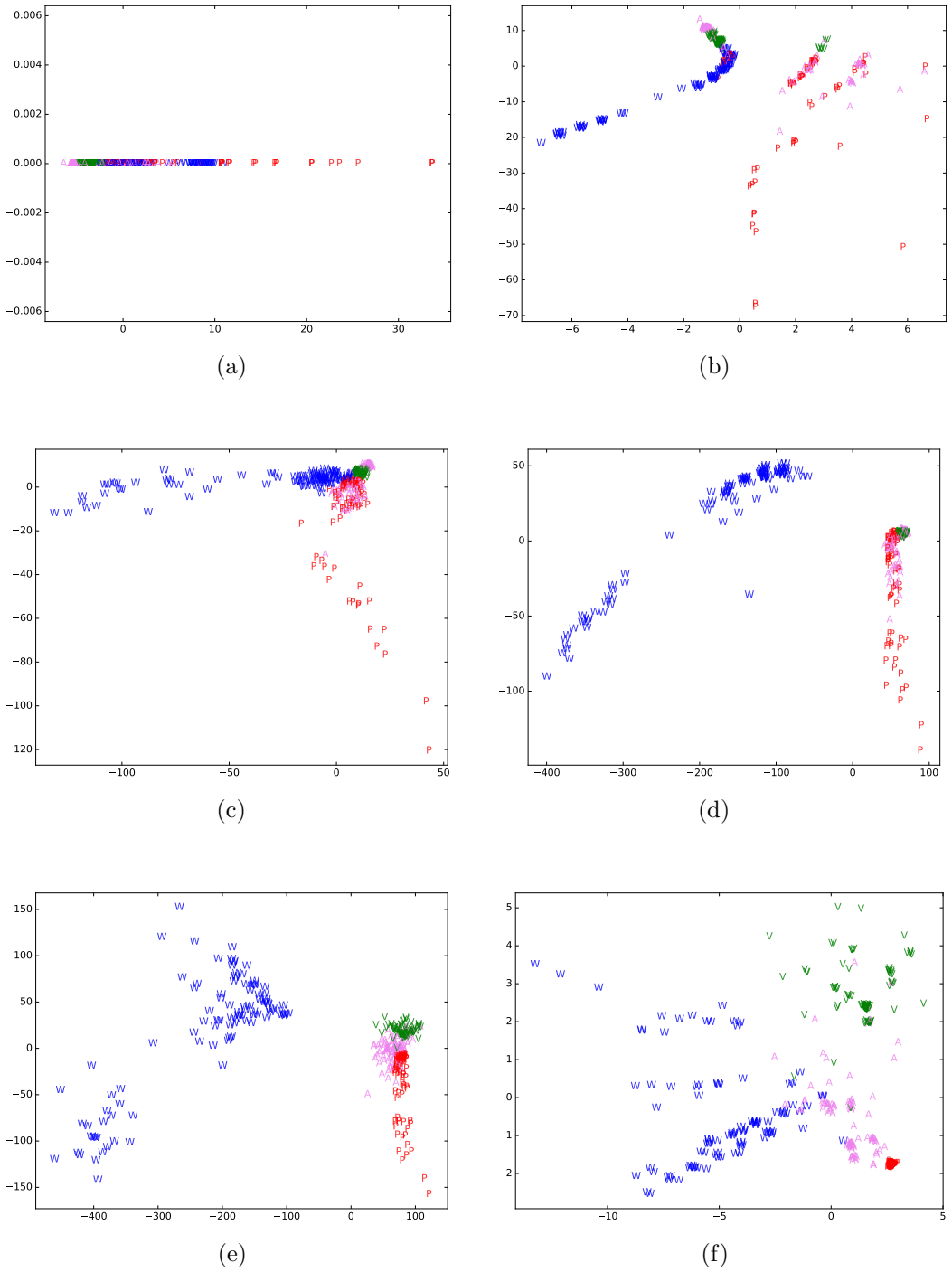


Figure 4: 2D projections of distance matrices for $4 \times 100$ evolved creatures and different sets of dissimilarity weights summarized in Table 2. The axes correspond to the two highest singular values of the centered distance matrix – they retain the maximum variance of the original distance matrix. Letters denote creatures with respect to the maximized fitness criterion used in evolution: red 'P' – height passive, violet 'A' – height active, green 'V' – velocity on land, blue 'W' – velocity in water. The fraction of total variance preserved in these projections is (a) 100%, (b) 97%, (c) 62%, (d) 93%, (e) 79% and (f) 62%. Agents in locations (d) are shown at the end of the video at `https://www.youtube.com/watch?v=lo4vL7gOuYk`.

cases. The distance matrices were then projected[1] into 2D, and the resulting projections are shown in Fig. 4. A small amount of random noise (jitter) was added to the coordinates of each point to avoid overplotting.

These projections can be analyzed from two points of view. First – as unlabeled data, when researchers look for clusters and investigate what such groups of similar organisms have in common. Second – as labeled data, where a researcher initially knows that they come from different sources and wants to discover and understand a potential relationship between these sources and similar traits in the organisms.

In the "unlabeled data" analysis, one investigates the shape of the distribution of data points ignoring colors and symbols – assuming that the origin of the agents is unknown. Fig. 4a is trivial because the similarity only reflects the number of vertices in creature bodies. We can see that for the majority of creatures, the number of vertices falls within some range, but there are also creatures with extraordinarily high or low number of vertices (the chart itself does not tell the number of vertices, it only indicates that these extreme agents are dissimilar from the majority). Fig. 4b represents similarity in the structure of bodies, and we can see a few clusters and rare outliers. We can only suppose that these clusters may have different origins. Figs. 4c and 4d illustrate geometrical similarity, with the latter chart considering the vertical axis as special in that the similarity measure was not allowed to tilt creatures when aligning them and assessing their geometrical similarity – it could only rotate them horizontally. One can see two stretched groups of creatures that share a common point in Fig. 4c, but not in Fig. 4d. This may indicate that some agents in these two groups share similar morphological characteristics, but they are rotated (agents from one group are more horizontal in shape, while agents from the other group are more vertical). Finally, Figs. 4d, 4e and 4f illustrate a transition from a pure geometrical similarity to the similarity in the number of neurons, with Fig. 4e being a mix of both aspects (geometry and neurons). We can see how considering the neural component (Fig. 4e) in addition to pure geometry (Fig. 4d) increased the resolution of the measure and allowed to better discriminate between geometrically identical agents. Another conclusion is that groups of agents based on the similarity in the distribution of neurons in the body (Fig. 4f) are different and less pronounced than clusters based purely on body geometry (Fig. 4d).

A separate set of conclusions can be drawn when one considers different origins of the agents (the "labeled data" analysis). In this particular experiment we not only know the origin (the four evolutionary goals), but we can also understand and interpret these goals, as they are meaningful for a human. Looking at Fig. 4a we can see that all agents evolved for speed on land ("V") are on the far end of the number of vertices in their bodies, but there are some "A" agents (they had maximized body height with neural network enabled) that are even more extreme in their number of vertices. Since Fig. 4a shows all the information from the distance matrix in one dimension, we conclude that this chart presents all agents sorted according to the number of vertices, and we can easily check that the left side of the plot corresponds to the simplest creatures, while the right side is occupied by the most complex agents – and these agents maximized their body height while being passive ("P"). Agents evolved for speed in water have an intermediate number of vertices in their bodies. Apart from very simple "A" agents, there is a separate subgroup of these agents with slightly higher body complexity. In Fig. 4b (similarity in body structure), we can see that swimmers ("W") constitute a separate group, but simple swimmers have similar structure to most runners ("V") and some jumpers ("A"). Tall structures ("P") are quite diverse, and again, the chart

[1]The distance matrix has been projected down by first computing the centering matrix for the distance matrix, then transforming the distance matrix into the cross-product matrix, and finally computing the singular value decomposition [23].

reveals two strategies that were evolved in the "A" group. Some "A" agents are structurally similar to runners (and in order to achieve high vertical position they perform jumping synchronized in time with the measurement period), and another subgroup of "A" agents does not use neural networks – their body structure evolved to be similar to "P" agents. The uniqueness of swimmers is confirmed in Figs. 4c and 4d, this time in terms of body geometry, and these two charts also show that geometrically, runners ("V") are extremely similar to each other. One swimmer is separate in Fig. 4d but not in Fig. 4c, and this is because its body, contrary to all other swimmers, is not horizontal when born under water. Finally, looking at neural similarity in Fig. 4f, we can see that this time all tall agents ("P") are identical because they hardly have any neurons (the neural network is not working in passive height maximization so there is no reason to develop it). The sequence of Figs. 4d–f – the progression from looking at pure body geometry to considering pure brain structure – demonstrates how the groups of runners, swimmers and jumpers expand and diversify while the group of passive tall agents collapses and becomes more uniform.

Given that there is not much computational cost required for such analysis, the presented technique provides an invaluable support tool for researchers analyzing large sets of evolved 3D agents. This particular experiment was just an illustration of a potential usage scenario; many fitness goals were mixed here and optimization was used to obtain a set of diverse morphologies, but the same approach could be used to investigate in detail agents evolved in one setup. The tool can be even more helpful when the origin of agents is unknown or cannot be easily labeled and interpreted (e.g. they come from an undirected, spontaneous evolution or coevolution).

### *3.3. Reconstructing sequences of mutations and evolutionary paths*

In paleontology and biostratigraphy, researchers often study traces of extinct organisms (such as exoskeletons of trilobites found in the fossil record), because the genomes of these organisms are not available. As they disappeared millions of years ago, classification of such organisms must be performed using phenotypic features such as selected aspects of morphology or physiology [24]. Another example of organisms with an extraordinary fossil record that are often used for testing paleoecological and evolutionary hypotheses are Foraminifera. Their bodies consist of connected chambers, and such a morphology can be represented as a 3D communication path (i.e., a linear graph) that connects successive chambers as they grow [25]. Even when the genomes of organisms are available, in many scenarios analyzing phenotypic traits may still be more practical (closer to the environment and fitness) than assessing similarity on a genetic level. Geometry (shape) of bodies is also crucial in most experiments in the areas of evolutionary design, collective robotics and evolutionary robotics.

The following experiment demonstrates the way the measure can be used to accurately estimate dissimilarity in order to discover ancestral relationships in a set of agents. Just as in the paleontological case described above, we assume that there is no information about any relationship between these agents. To generate sample morphologies, a number of consecutive mutations have been performed starting from a given initial genotype:

$$initial\ genotype \rightarrow mutant_1 \rightarrow mutant_2 \rightarrow \cdots \rightarrow mutant_N$$

The genetic encoding used in this experiment is called *direct* or *f0* [3, 15, 16]. In this encoding, the genotype explicitly lists coordinates of all the vertices. Each mutation moves a randomly selected vertex in an agent by a random distance in a random direction in 3D. The designs only differ in geometry, which makes it a demanding task for the dissimilarity measure, as no structural differences occur in the sequence. The initial genotype describes the *Humanoid* design – a simple structure with 16 vertices and 16 edges which resembles a

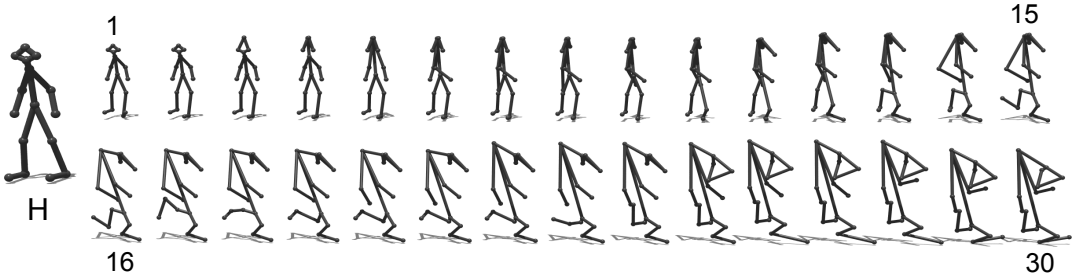


Figure 5: Sample mutation sequence. 30 consecutive geometrical mutations (top left to bottom right) of the initial *Humanoid* genotype.

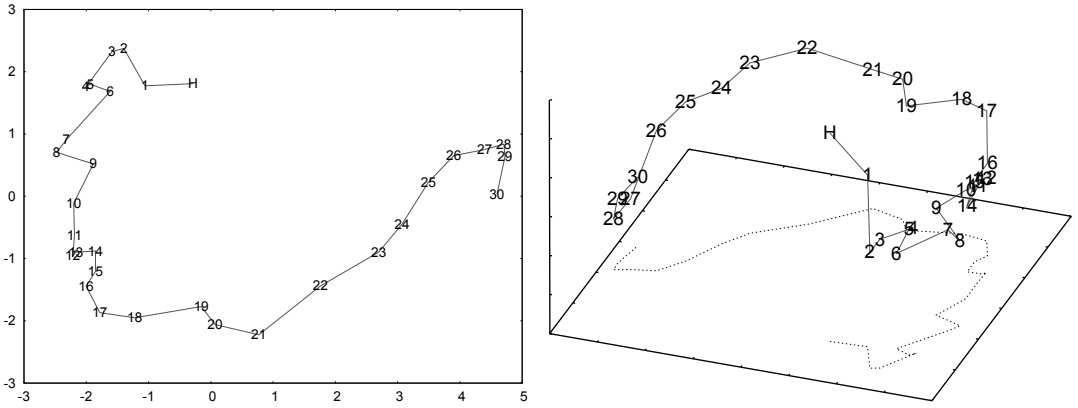

Figure 6: 2D and 3D projections of the distance matrix for the mutation sequence of the *Humanoid* genotype denoted as 'H'. The axes correspond to the two or three highest singular values of the centered distance matrix – they retain the maximum variance of the original distance matrix. The proportion of total variance preserved is 67% for 2D and 74% for 3D.

human body. Starting from this genotype, $N = 30$ mutations have been performed, as shown in Fig. 5.

The resulting 31 genotypes have been compared using the dissimilarity measure; a square matrix of size 31 has been constructed which contains distances between each pair of structures. As in the previous subsection, this matrix has been further processed using multidimensional scaling techniques to reduce it to two or three dimensions for visualization.

This experiment was repeated 10 times: 10 mutation sequences (each containing 31 genotypes) were generated independently, starting from the initial genotype. Fig. 6 shows the 2D and 3D plots for the experiment, which allowed the preservation of the highest amount of variation after reducing dimensionality. For all 2D projections, the proportion of total variance preserved ranged from 50% to 67%, and representative results of other runs are depicted in Fig. 7.

It can be clearly seen that the dissimilarity measure was able to capture the amounts of geometrical differences between mutated structures, which resulted in distinct paths visible in the graphs. Note that the reduction of dimensions causes loss of information, so that the patterns which can be seen in 2D and 3D graphs do not perfectly represent the original distance matrix. Still, the arrangement of mutants which results from dissimilarity estimation reflects the mutation process which occurred.

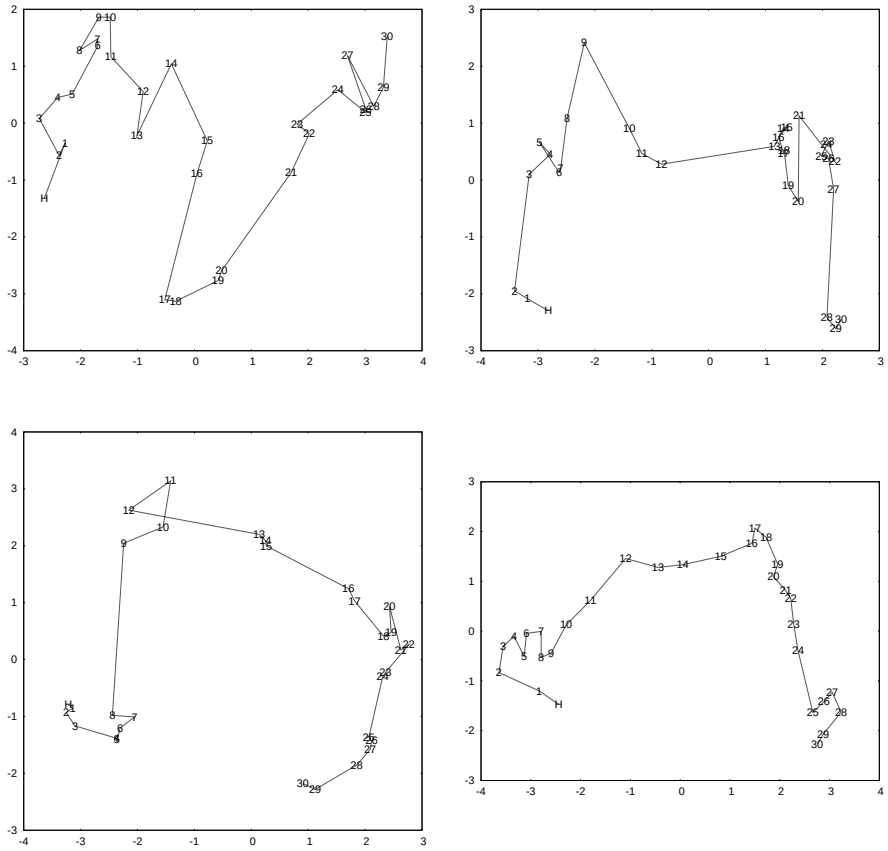

Figure 7: 2D projections of distance matrices for another four mutation sequences of the *Humanoid* genotype denoted as 'H'. The fraction of total variance preserved is (left to right, top to bottom): 58%, 49%, 57% and 65%. The bottom right graph corresponds to the sequence shown in Fig. 5.

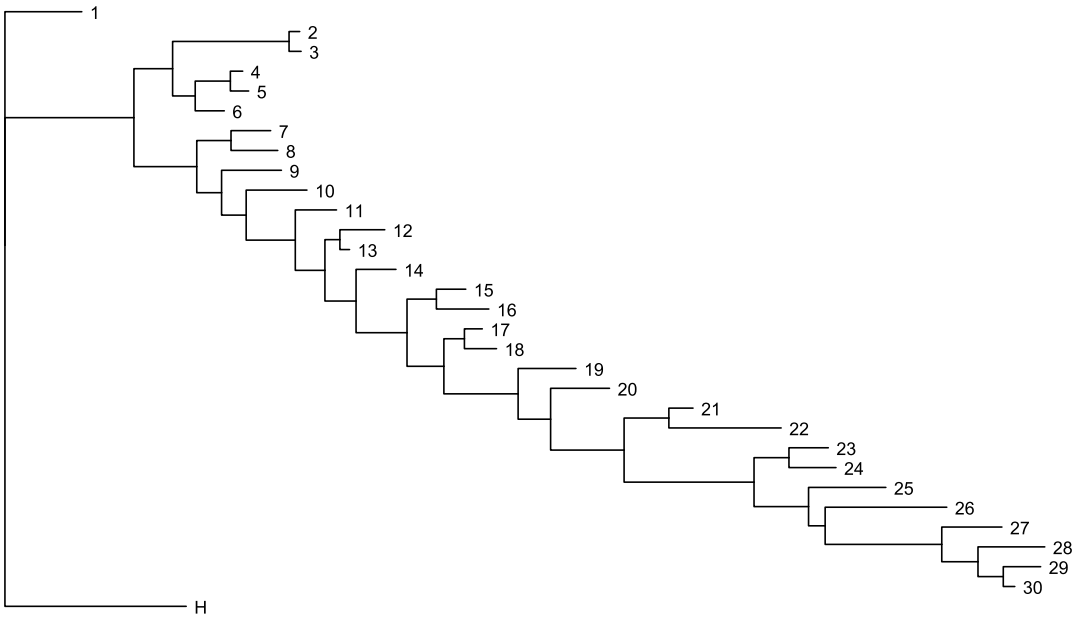


Figure 8: A tree produced by the neighbor-joining algorithm for the distance matrix also used in Fig. 6.

Fig. 8 shows a tree constructed using the heuristic neighbor-joining technique [26] from the distance matrix illustrated in Fig. 6. This unrooted tree is quite a successful reconstruction of the original sequence of mutations, and one can see analogies between the proximity of mutated designs in both figures. Reducing dimensions with the SVD and generating trees with the neighbor-joining algorithm yield two corresponding views of the same source data – the phenotypic distance matrix.

Now let us consider a more complex scenario. As in the previous experiment, a sequence of random geometric mutations is given, but this time the sequence bifurcates in the middle, and from this point on there are two independent mutation sequences. This is analogous to speciation:

$$\begin{aligned} \textit{initial genotype} \rightarrow \textit{mutant}_1 \rightarrow \cdots \rightarrow \textit{mutant}_{\frac{N}{2}} \rightarrow \textit{mutant}_{\frac{N}{2}+1} \rightarrow \cdots \rightarrow \textit{mutant}_N \\ \searrow \textit{mutant}^*_{\frac{N}{2}+1} \rightarrow \cdots \rightarrow \textit{mutant}^*_N \end{aligned}$$

A star (*) is used to denote one branch of the bifurcated mutation sequence. This experiment is illustrated in Fig. 9 for $N = 30$.

In each of the three subsequences (or two longer, overlapping subsequences), consecutive genotypes differ by a single geometrical mutation. For all 46 structures, a square matrix has been constructed which comprises distances between each pair. The matrix has been processed using the SVD transform [23] to obtain a two-dimensional projection for analysis. This experiment was repeated 10 times: 10 bifurcated mutation sequences (each containing 46 genotypes) were generated independently, starting from the initial genotype.

Representative results are shown in Fig. 10. Obviously, it is more difficult to distinguish between two independent mutation sequences, as they may become similar at any point. However, the similarity measure was still able to quantitatively capture geometrical differences between paths of mutated designs. Despite the information loss caused by the reduction of dimensions, trajectories of successive mutants can be seen in the 2D plots. Where the two

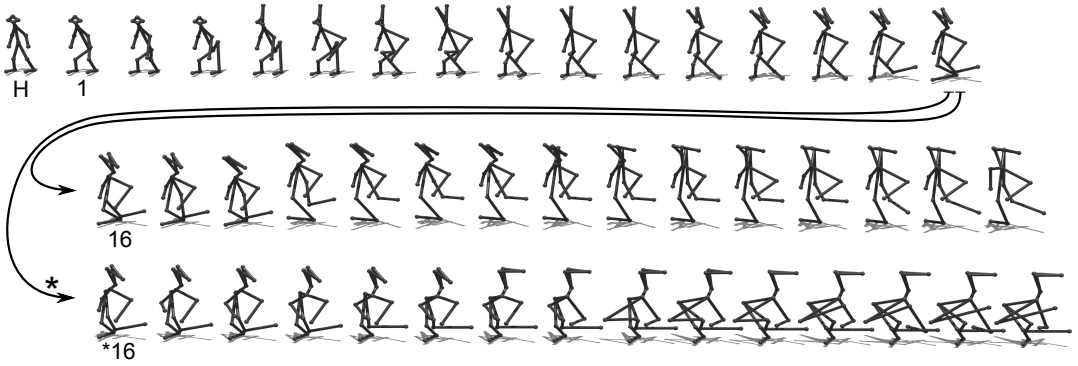


Figure 9: A bifurcated sequence of 30 consecutive geometrical mutations (left to right) of the initial *Humanoid* genotype, H.

paths are separate and clean, the corresponding trees produced by the neighbor-joining algorithm [26] also clearly show the bifurcation. Where mutation paths induced by phenetic dissimilarity overlap or cross, this is also reflected in the 2D graphs, and generated trees are unable to perfectly reconstruct original sequences of mutations.

### *3.4. Exploiting the relationship between phenetic properties and fitness criteria*

An often encountered need in research involving multiple evolved agents concerns identifying groups of similar agents, relating them with their environments, and selecting a small subset of representative agents for detailed inspection. Again, these steps could be done manually, but this would be extremely time-consuming, subjective, and impractical for large sets of agents – hence the need to develop automated methods. This issue has been addressed here by using the dissimilarity measure to identify fitness criteria and to automatically select representative designs by clustering evolved agents. The designs only differ in geometry – all of them share the same structure. Identical structures and the absence of control systems make the $w_V$, $w_D$ and $w_N$ weights irrelevant. The $w_G$ weight should be set to a positive value ($w_G = 1$ was used in this experiment).

In the evolutionary processes, five fitness criteria were employed:

- C: height of the center of mass,
- T: height of the top vertex of a design (i.e., maximal height),
- V: volume of a design (measured as its $size_x \cdot size_y \cdot size_z$),
- F: flatness (planarity) of a design (measured as $\frac{size_x \cdot size_y}{size_z}$),
- L: linearity of a design (along the $x$ axis, measured as $\frac{size_x}{size_y \cdot size_z}$),

where *size* is the extent of the bounding box along the given axis. These criteria model diverse environmental pressures: to keep most parts of a body elevated (C), to have the top part of a body as high as possible (T), to occupy the most space (V), to occupy the most area and stay flat (F), and to have the longest narrow body (L).

The initial design for these experiments is shown in Fig. 11. It is a very simple structure with 11 vertices and 10 edges of length 1. The simplicity of this design makes experimentation and analysis easier, yet the structure has the potential to form diverse geometric configurations.

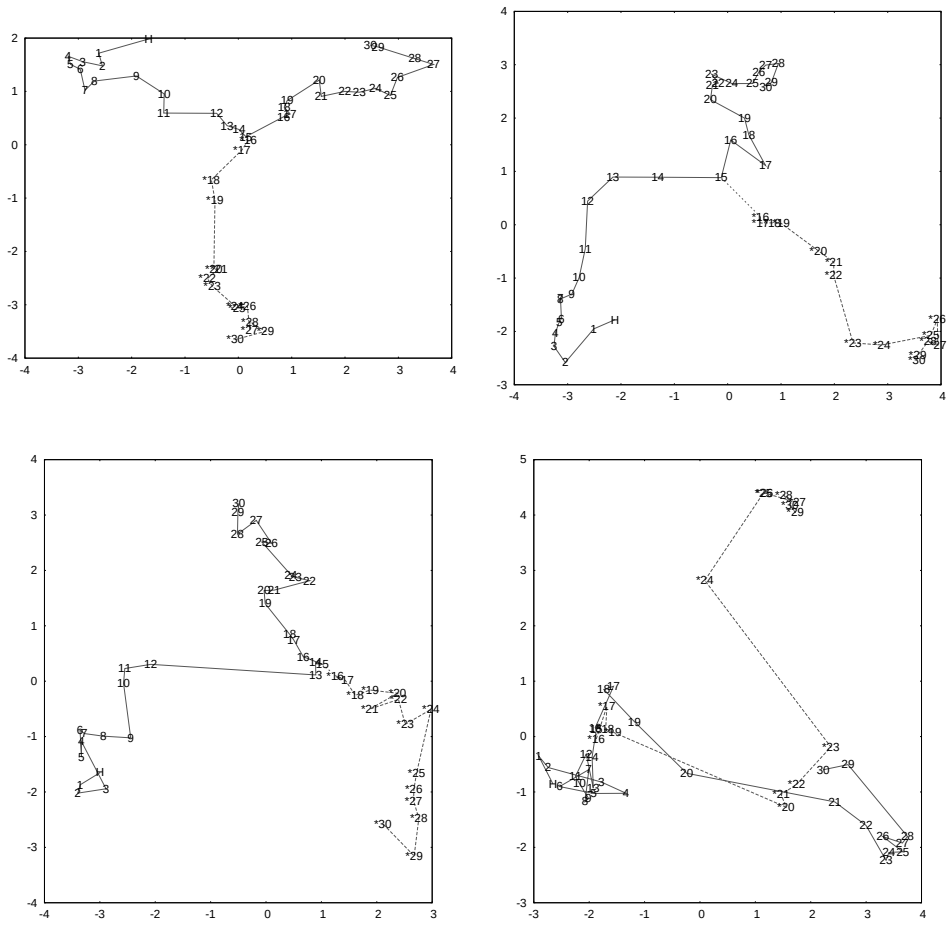

Figure 10: 2D projections of distance matrices for four bifurcated mutation sequences of the *Humanoid* genotype denoted as 'H'. The fraction of total variance preserved is (top to bottom, left to right): 54%, 56%, 47% and 50%. The bottom right graph represents the sequence shown in Fig. 9; compare how 29 and 30 became similar to *22 and *23, and notice the mutation of *23→*24 (elongation of the "right arm").

Figure 11: Initial structure for experiments with the evolution of geometry and five fitness criteria (Sect. 3.4).

For each of the five fitness criteria, 10 independent runs of a steady-state evolutionary algorithm were performed. The population size was 50. In order to preserve the structure, no crossing over was allowed, and only geometric mutations were used. The mutation randomly moved a random vertex of a body, and the maximal allowed edge length was 2. Tournament selection was used with a tournament size of 2. The evolutionary process consisted of two phases, each of which lasted until 1000 non-improving individuals were evaluated. In the first phase, negative selection consisted of removing a random genotype. In the second, high convergence "local search" phase, the worst genotype was removed by negative selection.

Although the analysis of parameters and dynamics of these 50 optimization processes is outside of the scope of this article, there are two issues worth mentioning:

- The "T" fitness criterion was not the one that yielded designs of maximal height. The highest designs (according to the "T" criterion) were evolved using the "C" goal, because its fitness landscape was easier for evolution and independently promoted all vertices (from base to top) to be moved upwards. The "T" goal only reflected the position of the topmost vertex, obscuring the influence of lower vertices on fitness.

- Among the ten "C"-evolved designs, there are three subgroups. In each case, the population was attracted to one of the three possible "ideas" or local optima, depending on which vertex was used as a base of the structure (see Fig. 11). The base vertex determines the maximal height that can be achieved. Once a particular base vertex dominated in the population, mutations improved the positions of remaining vertices and the population converged. Out of ten runs, there were six structures with one of the three outer leftmost or three outer rightmost vertices (Fig. 11) at the base, which allowed a maximum of four elevated edges (shown as C4 in Fig. 16). There was also one structure with the central vertex as the base (at most two elevated edges), and three structures with one of the remaining vertices at the base that allowed for at most three elevated edges (shown as C9 in Fig. 16).

Geometric dissimilarities among 50 resulting designs were computed, producing a distance matrix of size 50. Fig. 12 shows the UPGMA tree computed from this matrix. It can be seen that the group of six "C" designs (C0, C3, C5, C6, C4, C8) is considered most similar – these are the bodies similar to C4 in Fig. 16. Another cluster is formed by C1, C9, C2 and C7, where C1 and C2 are similar to C9 in Fig. 16, and C7 is the single structure with the central vertex at the base. As with all other "C" structures, C7 has all its edges elevated upwards, but it is the shortest of all the "C" designs; it is shown as the top right "C" in Fig. 13 and is located in the far distance on the right in Fig. 14.

Other bodies are considered less similar, yet they form relatively pure clusters. The "L" cluster is entirely separated, and it is further linked to the cluster of six elongated "C" structures. The "V" (volume geometries) cluster consists of nine "V"-evolved structures, while the "F" and "T" groups of structures are partially mixed.

The same distance matrix was also processed with the SVD to produce a 2D projection shown in Fig. 13. The relationships discussed above for the UPGMA clustering are also visible here: there is a clear "L" cluster, and two or three subgroups within the separated "C" cluster. Even though the proportion of total variance preserved is only 39%, one can easily distinguish generally unmixed "V", "F" and "T" clusters.

Fig. 14 depicts the actual designs located on a 2D plane at coordinates computed using SVD (shown in Fig. 13). This figure enables the drawing of a number of interesting conclusions on human versus automated ways of comparing designs. The first impression is that designs that are geometrically similar are located close to each other – the measure was able to capture the degree of visual similarity among 50 structures. Bodies in the "V" (maximized volume) group seem so complex that it would be extremely difficult to compare

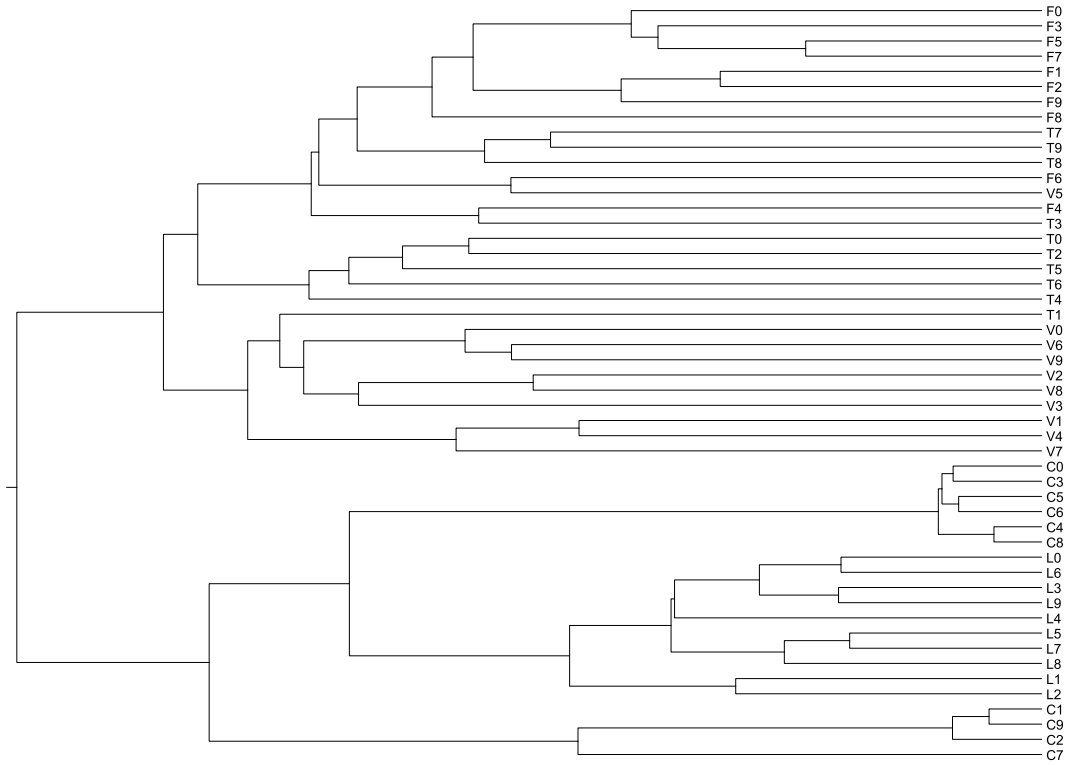


Figure 12: UPGMA clustering of 50 evolved agents (identical structure, 5 diverse fitness criteria).

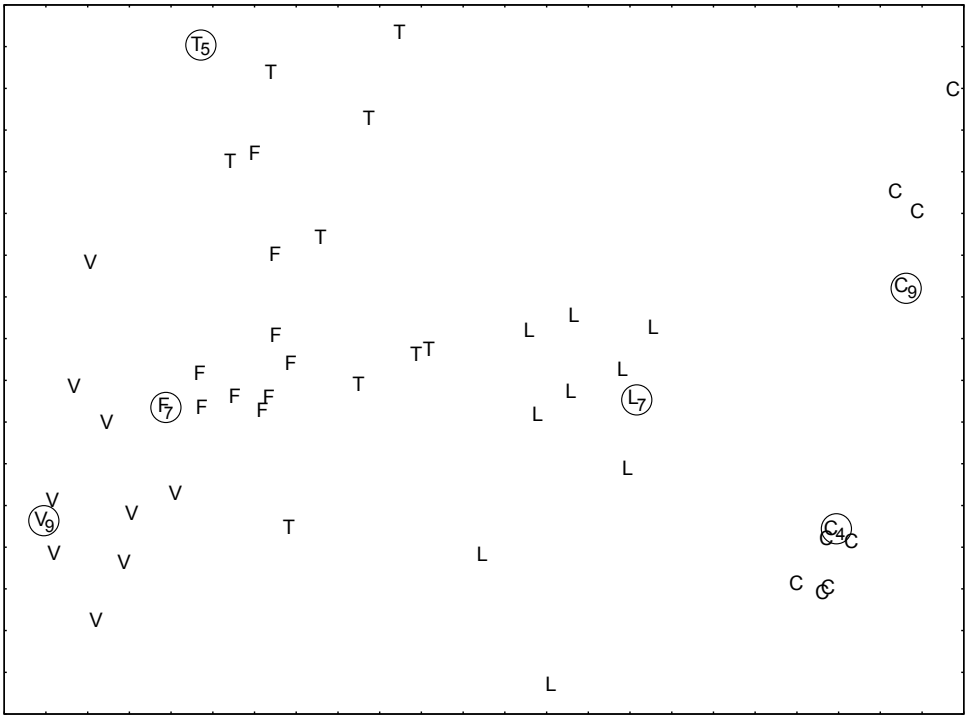


Figure 13: 2D SVD projection of the distance matrix of 50 evolved agents (identical structure, 5 diverse fitness criteria). The total variance preserved is 39%. Representatives are circled; see text for details.

them by hand. This is also true for the "T" (maximized height of the top vertex) group. It would be very hard for a human to classify evolved designs into either "V" or "T" which demonstrates the usefulness of the automated approach.

For a human, there are two unusual things about Fig. 14. Why is the "F" cluster of flat designs located between "V" and "T"? And why are the "L" designs so dispersed within their own cluster? There is one answer to both questions. While humans, by default, pay special attention to the vertical direction (or they are aware of its importance in this experiment), the dissimilarity estimation algorithm which has been employed in this experiment does not. The three axes are indifferent here, and there is no reference to any absolute coordinate system. The algorithm considers designs as located arbitrarily in a 3D space, so there is not much difference between the "L" and "C" groups. Since many of the "T" bodies are almost flat, it is wholly rational to consider them similar to the "F" structures.

On further analysis of individual designs, the layout proves to be most reasonable, with a general gradient from voluminous forms (left side) through flat (middle) to linear ones (right side). This ordering constitutes a direction in the phenetic space – a direction that reflects different concepts of fitness. It also displays the relative location of the areas in the phenetic space where locally optimal solutions exist for different fitness criteria.

One of the advantages of automated dissimilarity estimation is the ability to easily compute distance matrices for large sets of designs. These matrices can be further used to automatically select a small number of representatives from original populations. This is again very useful for a human, as it removes the need for manual and subjective inspection of optimized agents. To gain the benefits of this process, the PAM (Partitioning Around Medoids) method [27] was applied which is a more robust version of the $K$-means algorithm.

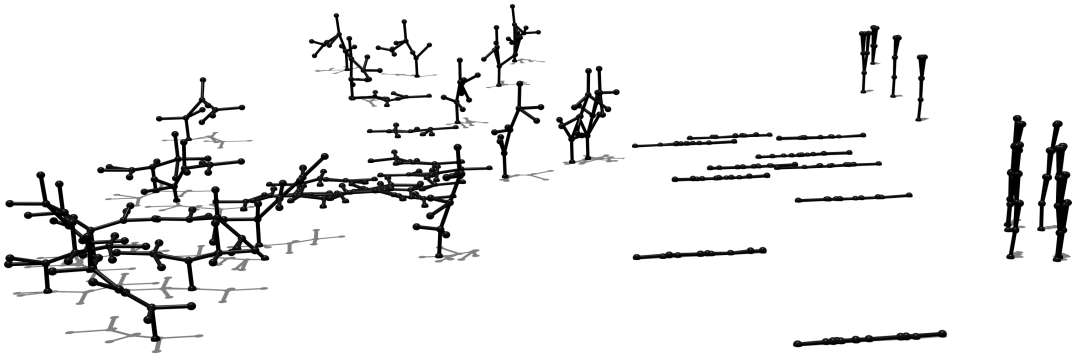

Figure 14: The actual bodies located on a 2D plane at coordinates from Fig. 13.

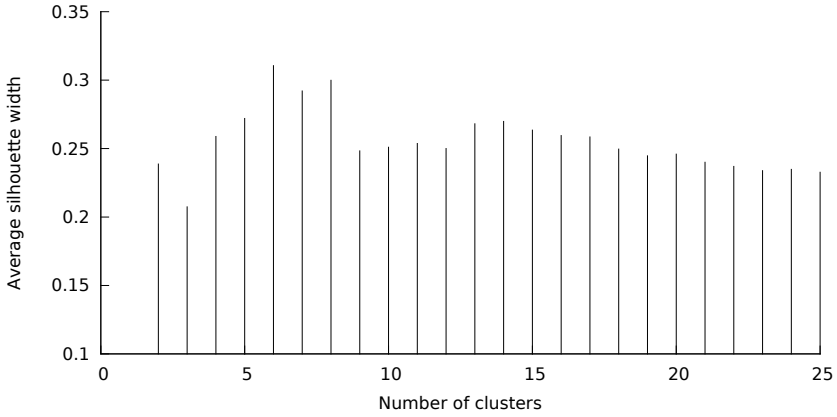


Figure 15: PAM clustering assessment.

In order to estimate the quality of clusters and automatically select the number of clusters, the average silhouette width was computed [28] for $K$ (the potential number of clusters) ranging from 3 to 50. The average silhouette width is a measure of separation of clusters, and for the 50 considered structures it has a maximum for $K = 6$, as shown in Fig. 15.

The PAM algorithm running for $K = 6$ produced the clusters shown below. Representative designs (medoids) are given first (before the colon), and they are also circled in Fig. 13.

$$
\begin{array}{lllllllllllll}
C_9 : C_1 & C_2 & C_7 & & & & & & & & & & \\
C_4 : C_0 & C_3 & C_5 & C_6 & C_8 & & & & & & & & \\
T_5 : T_0 & T_2 & T_3 & T_4 & T_6 & & & & & & & & \\
V_9 : V_0 & V_2 & V_3 & V_4 & V_5 & V_6 & V_7 & & & & & & \\
F_7 : F_0 & F_1 & F_2 & F_3 & F_4 & F_5 & F_6 & F_8 & F_9 & T_8 & V_2 & V_8 & \\
L_7 : L_0 & L_1 & L_2 & L_3 & L_4 & L_5 & L_6 & L_8 & L_9 & T_1 & T_7 & T_9 &
\end{array}
$$

The representatives which have been determined with no human intervention correspond to the five evolutionary goals, with the "C" goal having two representatives. The clusters are almost perfect and no random noise can be seen: the "C" clusters are clean, and the

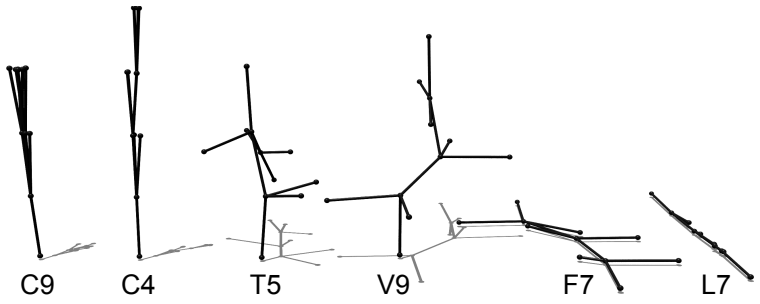


Figure 16: Six representatives of a set of 50 bodies evolved for 5 fitness goals.

"F" and "L" clusters contain all designs from their own fitness criterion and include a few similar designs from the other evolutionary goals. The "T" and "V" clusters are clean but incomplete – they miss 4 and 2 agents, respectively; the geometry of these "T" and "V" designs was similar enough to the "F" and "L" designs to assign them to clusters where "F" and "L" dominated. Overall, this result is extremely satisfactory given that the dissimilarity estimation was only concerned with the geometry of structures and not their horizontal/vertical orientation, and that the process of identification of clusters was fully automated and unaware of the origin of these 50 designs.

Fig. 16 depicts the six actual representatives. These pictures demonstrate that although it may be tempting for a human to intuitively estimate similarities, such a manual estimation would be far from perfect. The complete 3D structure is not perceived, and it is difficult for a human to rotate, reflect and move 3D agents to match corresponding vertices and objectively estimate their differences.

Increasing the size of the population (e.g. to 500 or 5000 individuals) would only use up additional time in order to compute the distance matrix (this can be easily parallelized), while the automated procedure to obtain representatives would stay the same and human effort would not increase, which was the goal.

## 4. Conclusions and further work

This article introduced and demonstrated applications of the similarity measure that support researchers in the analysis and organization of populations of three-dimensional agents and designs. Applications of this similarity evaluation procedure are independent from particular genetic encoding and genetic operators: measuring dissimilarity was performed on phenotypes, i.e., the actual structures, and not on their representations. This allowed the measure to be a more general tool, and at the same time moved the concept of similarity towards the quality and fitness of designs (i.e., from the genotype space to the phenotype space).

The advantages of using a similarity measure which have been presented in this article concern environments and systems where 3D agents are involved, such as evolutionary design, evolutionary robotics and artificial life. However, few assumptions have been made regarding compared graphs, so the 3D model is also compatible with biological forms, chemical and biochemical molecules, or more generally, structures which are built from a finite number of connected simple elements. These basic elements (vertices and edges) can be assigned various properties depending on the area of application – the properties can be masses, frictional coefficients, material types, strength of bonds, the number and type of attached control units, sensors and actuators, or their parameters.

An important area that highly benefits from analyses of genotype–phenotype relationships based on the similarity of phenotypes is optimization – in particular, the efficiency of optimization algorithms and the characteristics of the search space. Due to page restrictions, these experiments will be described separately.

In this work, estimating dissimilarity concerned static morphologies. Were they sequences of changing morphologies, there would be another dimension which characterizes such structures – time. Obviously, it is more difficult to estimate the similarity of moving structures, not only because of the computational complexity, but also because the similarity itself is not clearly defined and multiple approaches exist. When bodies change shape over time, it is their geometry that changes, but usually control signals change in time too. The similarity of two active agents – robots or animals, for example – can be based on both of these aspects: dynamics of control and sequences of movement. Both aspects are currently investigated in a long quest to develop a range of automated tools that support the analysis of life forms and their models.

**Acknowledgments:** This work has been supported by the Polish National Science Center, grant no. N N519 441939.

[1] V. Toğan, A. Daloğlu, Optimization of 3D trusses with adaptive approach in genetic algorithms, Engineering structures 28 (7) (2006) 1019–1027.

[2] P. Bentley, Evolutionary Design by Computers, Morgan Kaufman Publishers, 1999.

[3] M. Komosinski, A. Rotaru-Varga, Comparison of different genotype encodings for simulated 3D agents, Artificial Life Journal 7 (4) (2001) 395–418.

[4] M. Komosinski, M. Kubiak, Quantitative measure of structural and geometric similarity of 3D morphologies, Complexity 16 (6) (2011) 40–52. doi:10.1002/cplx.20367.
URL http://dx.doi.org/10.1002/cplx.20367

[5] H. Lipson, J. B. Pollack, Automatic design and manufacture of robotic lifeforms, Nature 406 (6799) (2000) 974–978.

[6] D. Marbach, A. J. Ijspeert, Co-evolution of configuration and control for homogenous modular robots, in: Proceedings of the Eighth Conference on Intelligent Autonomous Systems (IAS8), 2004, pp. 712–719.

[7] A. J. Ijspeert, A. Crespi, J. M. Cabelguen, Simulation and robotics studies of salamander locomotion. Applying neurobiological principles to the control of locomotion in robots, Neuroinformatics 3 (3) (2005) 171–196.

[8] V. G. Jáuregui, Tensegrity structures and their application to architecture, PUbliCan – Ediciones Universidad de Cantabria, 2010.

[9] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, K. Ramani, Three-dimensional shape searching: state-of-the-art review and future trends, Computer-Aided Design 37 (5) (2005) 509–530.

[10] D. Barthel, J. Hirst, J. Blazewicz, E. Burke, N. Krasnogor, ProCKSI: a decision support system for protein (structure) comparison, knowledge, similarity and information, BMC Bioinformatics 8 (1) (2007) 416–438.
URL http://www.biomedcentral.com/1471-2105/8/416

[11] P. J. Neal, L. G. Shapiro, C. Rosse, The digital anatomist structural abstraction: a scheme for the spatial description of anatomical entities, Journal of American Medical Informatics Association AMIA 98 (1998) 423–427.

[12] J. Foley, Computer Graphics: Principles and Practice, Addison-Wesley, Boston, 1997, Ch. 11.

[13] G. S. Hornby, J. B. Pollack, The advantages of generative grammatical encodings for physical design, in: Proceedings of the 2001 Congress on Evolutionary Computation, IEEE Press, 2001, pp. 600–607.

[14] M. Komosinski, A. Adamatzky (Eds.), Artificial Life Models in Software, 2nd Edition, Springer, London, 2009. doi:10.1007/978-1-84882-285-6.
URL http://www.springer.com/978-1-84882-284-9

[15] M. Komosinski, S. Ulatowski, Framsticks: Creating and understanding complexity of life, in: M. Komosinski, A. Adamatzky (Eds.), Artificial Life Models in Software, 2nd Edition, Springer, London, 2009, Ch. 5, pp. 107–148.

[16] M. Komosinski, S. Ulatowski, Framsticks web site, http://www.framsticks.com (2016).

[17] M. Komosinski, S. Ulatowski, Genetic mappings in artificial genomes, Theory in Biosciences 123 (2) (2004) 125–137. doi:10.1016/j.thbio.2004.04.002.
URL http://dx.doi.org/10.1016/j.thbio.2004.04.002

[18] M. Hapke, M. Komosinski, Evolutionary design of interpretable fuzzy controllers, Foundations of Computing and Decision Sciences 33 (4) (2008) 351–367.
URL http://www.framsticks.com/files/common/Komosinski_EvolveInterpretableFuzzy.pdf

[19] W. Jaskowski, M. Komosinski, The numerical measure of symmetry for 3D stick creatures, Artificial Life Journal 14 (4) (2008) 425–443. doi:10.1162/artl.2008.14.4.14402.
URL http://dx.doi.org/10.1162/artl.2008.14.4.14402

[20] M. Komosinski, S. Ulatowski, Framsticks SDK (Software Development Kit).
URL http://www.framsticks.com/sdk

[21] M. Komosinski, G. Koczyk, M. Kubiak, On estimating similarity of artificial and real organisms, Theory in Biosciences 120 (3-4) (2001) 271–286. doi:10.1007/s12064-001-0023-y.
URL http://dx.doi.org/10.1007/s12064-001-0023-y

[22] P. H. A. Sneath, R. R. Sokal, Numerical Taxonomy: The Principles and Practice of Numerical Classification, W. H. Freeman and Company, San Francisco, 1973.

[23] C. D. Meyer, Matrix Analysis and Applied Linear Algebra, Society for Industrial and Applied Mathematics (SIAM), 2000.

[24] R. A. Fortey, Trilobite systematics: the last 75 years, Journal of Paleontology 75 (6) (2001) 1141–1151.

[25] M. Komosinski, A. Mensfelt, P. Topa, J. Tyszka, Application of a morphological similarity measure to the analysis of shell morphogenesis in Foraminifera, in: A. Gruca, A. Brachman, S. Kozielski, T. Czachórski (Eds.), Man–Machine Interactions 4, Vol.

391 of Advances in Intelligent Systems and Computing, Springer, 2016, pp. 215–224. doi:10.1007/978-3-319-23437-3_18.
URL http://dx.doi.org/10.1007/978-3-319-23437-3_18

[26] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular Biology and Evolution 4 (4) (1987) 406–425.

[27] L. Kaufman, P. J. Rousseeuw, Finding groups in data. An introduction to cluster analysis, Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Wiley, New York, 1990, Ch. 2.

[28] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1) (1987) 53–65.