DOI: http://dx.doi.org/10.7551/978-0-262-32621-6-ch056

# An informational study of the evolution of codes in different population structures

Andrés C. Burgos[1] and Daniel Polani[2]

[1,2]Adaptive Systems Research Group, University of Hertfordshire, Hatfield, UK
[1]a.c.burgos@herts.ac.uk



**Abstract**

We consider the problem of the evolution of a code within a structured population of agents. The agents try to maximise their information about their environment by acquiring information from the outputs of other agents in the population. A naive use of information-theoretic methods would assume that every agent knows how to "interpret" the information offered by other agents. However, this assumes that one "knows" which other agents one observes, and thus which code they use. In our model, however, we wish to preclude that: it is not clear which other agents an agent is observing, and the resulting usable information is therefore influenced by the universality of the code used and by which agents an agent is "listening" to.

## Introduction

If we consider organisms capable of processing information, then we can argue that they must be able to internally assign meaning to the symbols they perceive in a code-based manner (Görlich et al., 2011). For instance, bacteria perceives chemical molecules in their environment and interprets them in order to better estimate environmental conditions and (stochastically) decide their phenotype (Platt and Fuqua, 2010). Plants detect airborne signals released by other plants, being able to interpret them as attacks of pathogens or herbivores (Heil and Karban, 2010). Therefore, a correspondence between environmental conditions and chemical molecules must be established. It is in this way that Barbieri characterises codes, and he proposes three fundamental characteristics for them: they connect two independent worlds; they add meaning to information; and they are community rules (Barbieri, 2003).

Codes connect two independent worlds by establishing a correspondence or mapping between them. These worlds are independent and thus there are no material constraints for establishing arbitrary mappings. The meaning of information comes exclusively from the mapping: symbols by themselves are meaningless. Finally, the third property requires that the correspondence between the two worlds constitutes an integrated system.

For instance, human languages establishes a correspondence between words and objects (Barbieri, 2003); in bacteria it is between chemical molecules and environmental conditions (Waters and Bassler, 2005). Words (or chemical molecules) by themselves do not have any meaning, they are chosen arbitrarily and any individual of a population can define its own set together with its mapping. However, populations of individuals sharing the same code are ubiquitous in nature. How is it that codes come to be shared by many individuals when, as stated, they are completely arbitrary? This question is what we are investigating in the present paper.

For this work, we assume a simple scenario where organisms seek to maximise their long-term growth rate by following a bet-hedging strategy (Seger and Brockmann, 1987). We know that maximising their information about the environment achieves this (Shannon, 1948; Kelly, 1956; Donaldson-Matasci et al., 2010). Then, individuals obtaining side environmental information from other individuals will have an advantage over those that do not, since they would be able to better predict the future environmental conditions. However, for individuals to be able to communicate with each other, they must be able to translate symbols into environmental conditions, where the

output of these symbols results from an individual's code. The code of an individual is a stochastic mapping from its sensors states to a set of outputs.

For this study, we consider outputs of individuals (or agents) as conventional signs. In semiotics, the science of all processes in which signs are originated, stored, communicated, and being effective (Görlich et al., 2011), two types of signs are traditionally recognised: *conventional signs* and *natural signs* (Deely, 2006). In conventional signs there is no physical constraint on the possible mappings, they are established by conventions. On the other hand, in natural signs, there is always a physical link between the signifier and signified, such as smoke as a sign of fire, odours as signs of food, etc. (Barbieri, 2008).

We are not interested in the particular detailed mechanisms by which an agent implements its code, nor how the agent decodes the outputs of other agents. Instead, we focus on the theoretical limits on the amount of environmental information an agent can possibly acquire resulting from different scenarios of population structure and codes distribution.

The natural framework to analyse such quantities is information theory (Shannon, 1948). However, it does not take semantic aspects into account, it only deals with frequencies of symbols instead of what they symbolise. Codes, on the other hand, add meaning to information, which makes the integration of sciences such as semiotics with information theory non-trivial (Favareau, 2007; Battail, 2009). In the following section, we present an information-theoretic model which incorporates the necessity of conventions by dropping from the model the usual implicit assumption of knowing the identity of the communicating units.

## Model

To introduce the model in a progressive manner, let us first consider three agents, $\theta_1$, $\theta_2$ and $\theta_3$. Each of these agents depend on the same environmental conditions for survival, which are modelled by a random variable $\mu$. Agents acquire information about the environment through their sensors, which are modelled by random variables $Y_{\theta_1}$, $Y_{\theta_2}$ and $Y_{\theta_3}$, all three conditioned on $\mu$, for agents $\theta_1$, $\theta_2$ and $\theta_3$, respectively. We assume each agent acquires the same amount and aspects of environmental information from $\mu$, *i.e.* $p(Y_{\theta_1}|\mu) = p(Y_{\theta_2}|\mu) = p(Y_{\theta_3}|\mu)$. Let us further assume that the information each agent acquires about the environment does not eliminate its uncertainty, *i.e.* $H(\mu|Y_{\theta_i}) > 0$ for $1 \leq i \leq 3$. The code of an agent is a stochastic mapping from its sensor states into a set of outputs, and is represented by the conditional probabilities $p(X_{\theta_1}|Y_{\theta_1})$, $p(X_{\theta_2}|Y_{\theta_2})$ and $p(X_{\theta_3}|Y_{\theta_3})$ for agents $\theta_1$, $\theta_2$ and $\theta_3$, respectively (see Fig. 1).

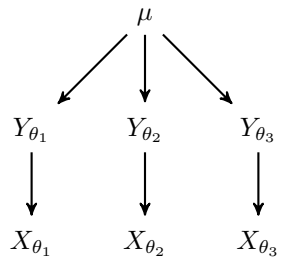


Figure 1: Bayesian network representing the relantionship between the sensor and output variables of three agents.

Let us assume that agent $\theta_1$ perceives only the outputs of agents $\theta_2$ and $\theta_3$. One possible way of computing the information about the environment agent $\theta_1$ has is to consider the mutual information between $\mu$ and the joint distribution of the sensor of $\theta_1$ and the outputs of $\theta_2$ and $\theta_3$: $I(\mu; Y_{\theta_1}, X_{\theta_2}, X_{\theta_3})$. However, by writing down this quantity, we are implicitly assuming that agent $\theta_1$ "knows" which output corresponds to $\theta_2$ and which output corresponds to $\theta_3$. Therefore, in this consideration, an agent can theoretically do the translations of the outputs according to some internal model of other agents and infer the mentioned amount of information about its environment.

### Indistinguishable sources

For this study, on the contrary, we consider an agent observing other agents' messages, but under the assumption that the originator of a message cannot be identified. In this way, the total amount of information an agent can infer from the outputs of other agents will depend on to which extent it either can identify who the other agents are or can rely on them using a coding scheme that does not depend too much on their particular identity. For instance, if agents $\theta_2$ and $\theta_3$ both

agree on the output for each of the environmental conditions, then agent $\theta_1$ should be able to infer more environmental information than if they disagree on the output for each of the environmental conditions, given that agent $\theta_1$ does not know which of the agents it is observing.

To model this idea, let us assume a random variable $\Theta'$ denoting the selected agent, which depends on the same environmental conditions for survival, which are modelled, as above, by a random variable $\mu$. Agents acquire information about the environment through their sensors, which are modelled by a random variable $Y_{\Theta'}$ conditioned on the index variable denoting the agent under consideration, $\Theta'$, and $\mu$. The amount of acquired sensory information of a specific agent $\theta'$ about $\mu$ is given by $I(\mu; Y_{\theta'})$. As above, the code of an agent is a stochastic mapping from its sensor states into a set of outputs, and is represented by the conditional probability $p(X_{\theta'}|Y_{\theta'})$ for an agent $\theta'$ (see Fig. 2).

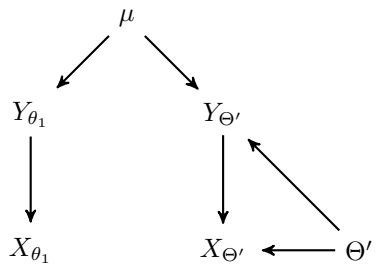


Figure 2: Bayesian network representing the relationships as described above (see text).

However, now we want to model the fact that we do not know which agent is observed. In the case with maximum uncertainty, $\Theta$ is uniformly distributed, and then this parametrisation of the codes considers the outputs of all agents in $\Theta'$ altogether, such that if we are not observing $\Theta'$, we cannot identify whose agent's output we are observing. In Eq. 3 and Eq. 4 we show two examples of codes for agents $\theta_2$ and $\theta_3$, while their sensor states are define by the Eq. 2 (Eq. 1 defines the sensors states of agent $\theta_1$). We compute how much information about the environment there is when the outputs of both agents ($\theta_2$ and $\theta_3$) are considered together by agent $\theta_1$.

$$Pr(Y_{\theta_1}|\mu) \coloneqq \begin{array}{c} \\ \mu_1 \\ \mu_2 \end{array} \begin{array}{c} \begin{array}{cc} y_1 & y_2 \end{array} \\ \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix} \end{array} \quad (1)$$

$$Pr(Y_{\Theta'}|\mu, \Theta') \coloneqq \begin{array}{c} \\ \theta_2,\ \mu_1 \\ \theta_2,\ \mu_2 \\ \theta_3,\ \mu_1 \\ \theta_3,\ \mu_2 \end{array} \begin{array}{c} \begin{array}{cc} y_1 & y_2 \end{array} \\ \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \\ 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix} \end{array} \quad (2)$$

$$Pr(X_{\Theta'}|Y_{\Theta'}, \Theta') \coloneqq \begin{array}{c} \\ \theta_2,\ y_1 \\ \theta_2,\ y_2 \\ \theta_3,\ y_1 \\ \theta_3,\ y_2 \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad (3)$$

$$Pr(X_{\Theta'}|Y_{\Theta'}, \Theta') \coloneqq \begin{array}{c} \\ \theta_2,\ y_1 \\ \theta_2,\ y_2 \\ \theta_3,\ y_1 \\ \theta_3,\ y_2 \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array} \quad (4)$$

If we assume $p(\theta_2) = p(\theta_3) = 1/2$, and $p(\mu_1) = p(\mu_2) = 1/2$, then if we consider the codes shown in Eq. 3, we have that $I(\mu; Y_{\theta_1}, X_{\Theta'}) = 0.97872$ bits, where $\Theta'$ consists of agents $\theta_2$ and $\theta_3$. However, had $\theta_2$ and $\theta_3$ "opposite" codes as shown in Eq. 4, then $I(\mu; Y_{\theta_1}, X_{\Theta'}) = 0.9192$ bits, which is exactly $I(\mu; Y_{\theta_1})$, that is, $I(\mu; X_{\Theta'}|Y_{\theta_1}) = 0$ bits (agent $\theta_1$ cannot acquire any side information from the outputs of agents $\theta_2$ and $\theta_3$). We should note here that the sensor states $y_1$ and $y_2$ of agents $\theta_2$ and $\theta_3$ in the conditional probability shown in Eq. 3 and 4 refer almost deterministically to the same environmental condition, and therefore the loss of side information is thus entirely due to the incompatible codes. The conditional probabilities of sensor states given the environmental conditions further defined throughout the paper are also assumed to be almost deterministically.

## Population information

The model shown in Fig. 2 considers the environmental information of agent $\theta_1$, ignoring its own output $X_{\theta_1}$. Nevertheless, agents ignoring their outputs is contrary to our assumption over the sources of the outputs. To incorporate this option in the model shown in Fig. 2, we could consider the state space of $\Theta'$ as the set $\{\theta_1, \theta_2, \theta_3\}$. Then, to express not only the environmental information of agent $\theta_1$, but the average environmental information of the whole population, we can parametrise the sensors of the agents by a random variable $\Theta$ (defined over the same state space, representing the same set of agents as $\Theta'$), such that

$p(Y_\Theta|\mu,\Theta) = p(Y_{\Theta'}|\mu,\Theta')$ (*i.e.*, $Y_{\Theta'}$ is *i.i.d.* to $Y_\Theta$, and vice versa).

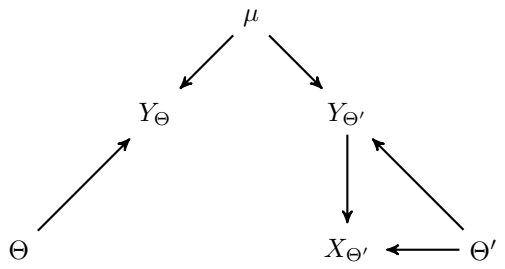


Figure 3: Bayesian network representing the sensor variables of a set of agents indexed by the random variable $\Theta$, and the sensor and output variables of a copy of the set of agents indexed by $\Theta$ named $\Theta'$.

In this way, the average environmental information of a population of the agents selected by $\Theta$ is given by $I(\mu; Y_\Theta, X_{\Theta'})$ (see Fig. 3). This measure can be consider as the objective function to maximise in our model. However, we would be making two important assumptions: first, this objective function assumes agents have access to the environmental conditions $\mu$, which they indirectly do but only through their sensors; and second, every agent would perceive the output of every other agent, including itself. In this work, we instead simplify the model in that we propose agents following a behaviour such that it maximises the similarity of their outputs (via their codes) with those of which the agent perceives. A consequence of this behaviour is that the average information about $\mu$ is also maximised. In addition, we will introduce a potentially flexible "population structure", so that we can specify which agents interact with which.

## Code similarity

First, we introduce a copy of the codes of the agents, such that when we instantiate the variables $X_\Theta$ and $X_{\Theta'}$, the probabilities are the same. The structure of the population is then given by $p(\Theta,\Theta') = p(\Theta)p(\Theta')$. However, the conditional independence of $\Theta$ and $\Theta'$ restricts significantly the diversity of the structures that can be represented. In order to model a general interaction structure between agents, we consider $p(\Theta,\Theta')$ not independent, as shown in the Bayesian network in Fig. 4, where we introduce a helper variable $\Xi$.

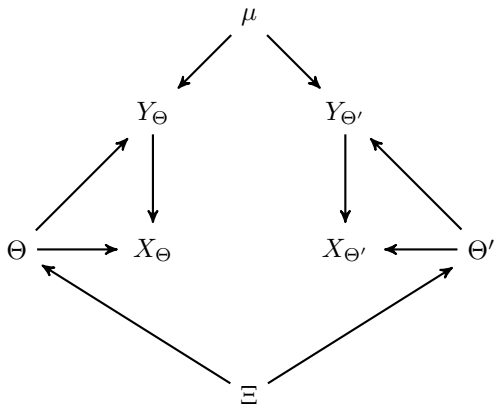


Figure 4: Bayesian network representing the relantionship of the variables in the model of code evolution. $Y_{\Theta'}$ is an *i.i.d* copy of $Y_\Theta$ and $X_{\Theta'}$ is an *i.i.d.* copy of $X_\Theta$. $\Theta'$ covers the same set of agents as $\Theta$, but its probability distribution is not necessary the same.

We can interpret our objective function $I(X_\Theta; X_{\Theta'})$ as the average code similarity of a population of agents according to the population structure $p(\Theta,\Theta')$. For instance, if two agents cannot exchange outputs in a given structure, then there is no gain (in the value of the objective function) in adopting similar codes for these two agents.

If we consider our system as a process in time, then at each time-step two agents are chosen according to $p(\Theta,\Theta')$. Agent $\Theta$ reads the output of agent $\Theta'$ (generated via its code, which is *i.i.d* over time), and let us assume that it stores the pair $(Y_\Theta, X_{\Theta'})$, *i.e.* its current sensor state together with the perceived output. If this is repeated a large number of times, then the total amount of environmental information that can be inferred from the collected statistics by the population is bounded by $I(\mu; Y_\Theta, X_{\Theta'})$. This is the theoretical limit to which we refer in the introduction, and for this study we are not interested in how the inference is computed. However, we implicitly assume that agents decode the perceived outputs according to their codes.

## Distance between two codes

In order to visualise the evolution of codes, we define the distance between the codes of two agents $\theta_i$ and $\theta_j$ as the square root of the *Jensen-Shannon divergence* (Wong and You, 1985; Lin, 1991) be-

tween them. This measure has the property that $0 \leq JSD(\theta_i, \theta_j) \leq 1$ when $\log_2$ is used, and the square root yields a metric. Let us note that this distance requires the sensor states $Y$ to be named identically (for the corresponding states of $\mu$) among agents in order to be meaningful. As we stated above, this is (closely) the case in all our experiments. This requirement over the sensor states discards the possibility of using other measures such as mutual information.

## Methods

To illustrate the behaviour of our model, we consider three different scenarios, which are described in the Results section. The common parameters for the first two experiments are the following: the population consists of 25 agents (the small number was chosen to avoid high computational costs); the amount and quality of the acquired sensory information is the same for every agent, that is $p(Y_{\theta_i}|\mu) = p(Y_{\theta_j}|\mu)$ for every $i, j \in [1, 25]$. For the last scenario, the only difference is that we consider only 15 agents.

The optimisation algorithm used in the following experiments is CMA-ES (Covariance Matrix Adaptation Evolution Strategy), which is a stochastic derivative-free method for non-linear optimisation problems (Hansen and Ostermeier, 2001). We utilised the implementation provided by the Shark library v3.0.0 (Igel et al., 2008) with its default parameters, which implements the CMA-ES algorithm described in (Hansen and Kern, 2004). The evolutionary algorithm used for optimisation does not intend to represent the actual evolution of the codes. Instead, we are interested in the solutions of this optimisation process, which are representative of the possible outcomes of evolution.

To visualise the evolution of the codes of the agents, we use the method of multidimensional scaling provided by R version 2.14.1 (2011-12-22).

## Results

In this section, we analyse the outcome of three different scenarios where code similarity is maximised. While the outcomes are particular for one simulation, they are illustrative of the richness that the model is able to capture, which is described for each scenario. The outcomes are typical solutions, and we cannot perform statistics over simulations since the many solutions are qualitatively different.

### Well-mixed population

In the first scenario, each agent $\theta_i$ perceives the output of every other possible agent $\theta_j$ with the same probability, that is $p(\theta_i, \theta_j) = 1/25^2$ for every $i, j \in [1, 25]$. The maximum average code similarity is bounded by $I(Y_\Theta; Y_{\Theta'}) = 1.71908$ bits, which is achieved under two conditions: first, every code must be a one-to-one mapping; second, the code must be universal. This is indeed the outcome of the performed optimisation, as we show in Fig. 5: the optimised codes (blue points) converged into a universal code (the distance between any of them is zero). Each red point correspond to an initial code.

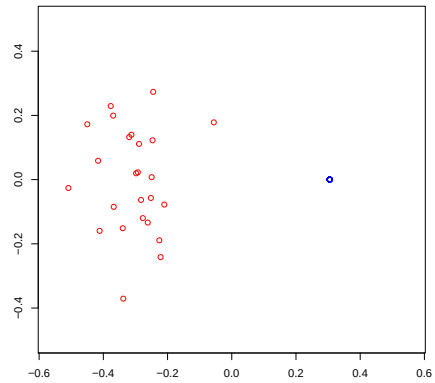


Figure 5: 2-dimensional plot of code distance: red points are codes at the beginning of the optimisation process; blue points are codes at the end of the optimisation process (where the distance between every pair of codes is zero).

The resulting code adopted by the population is a one-to-one mapping between sensor states and outputs, and any of the 24 possible one-to-one mappings is a global maximum (there are 4 sensor states and 4 possible outputs). However, it is still interesting to briefly analyse the possible paths towards a universal and optimal code. In Fig. 6, we show the distribution of the adopted codes by the agents of the population in a moment of the optimisation process where the average code similarity is $I(X_\Theta; X_{\Theta'}) = 1.18276$ bits. Here, the most popular code is the suboptimal code shown

in Fig. 6 (a). This results from the particular initialised codes, driving the agents temporarily towards a suboptimal code. However, once any of the many-to-one codes becomes (nearly) universally adopted, then any code's deviation improving code similarity will eventually drive the convention towards optimality. The fact that it does not need simultaneous changes in the code increases the likeliness of improving the code similarity.

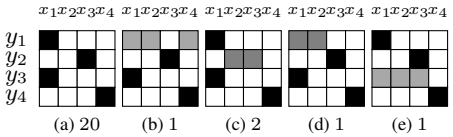


Figure 6: Representation of the codes $p(x|y)$ by a heat-map using inverse grayscale. For each evolved code, we output the number of agents adopting it. This code distribution was achieved with 25 agents in a well-mixed population.

## Spatially-structured population

In another set-up, we assume the agents are structured in a $5 \times 5$ grid, where $p(\theta, \theta') = 1/105$ if $\theta$ and $\theta'$ are neighbours or when $\theta = \theta'$. After randomly initialising the codes, the performed optimisation plateaued on an average code similarity of $I(X_\Theta; X_{\Theta'}) = 1.13536$ bits. As in the former scenario, here the optimal solution is also a universal code with a one-to-one mapping. However, in this case, the result is not a universal code, as can be appreciated in Fig. 7. Spatially structure populations are sensitive to the initial codes and how codes are updated.

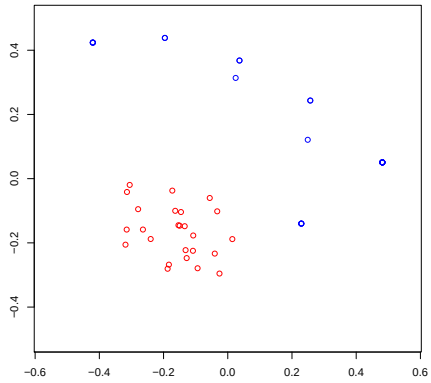


Figure 7: 2-dimensional plot of code distance: red points are codes at the beginning of the optimisation process; blue points are codes at the end of the optimisation process.

The resulting code distribution among the population is shown in Fig. 8, with 8 different codes in the population. Different from the well-mixed population structure, in a spatially structured population the pressure to agree on a code occurs only between neighbours. A consequence of this is that many local conventions are established within neighbourhoods, and once this situation is reached, to improve the total code similarity, some simultaneous changes to the agent's codes would be needed. For instance, the code shown in Fig. 8 (e) could increase the average similarity of the population if $p(x_2|y_1) = 1$, as it is in the rest of the codes. However, for this to happen (in this particular case), at least two agents need to change their code simultaneously (otherwise the average similarity decreases), which makes the deviation from the resulting code distribution unlikely.

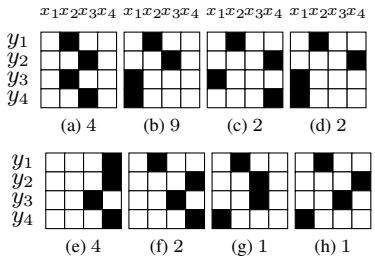


Figure 8: Representation of the codes $p(x|y)$ by a heat-map using inverse grayscale. For each evolved code, we output the number of agents adopting it. This code distribution was achieved with 25 agents in a grid structure.

## Free structure

For our last scenario, we let the structure evolve with the codes without any constraint. In this case, the resulting average code similarity is nearly optimal, but the code is not necessarily universal. This is because, when the structure is not fixed, agents form roughly disconnected clusters of related codes. In this process, the interaction probability of agents with unrelated codes will vanish, decreasing the entropy of the population structure $H(\Xi)$ (see Fig. 9). However, once the clusters are formed, if it is not a single isolated agent (such that nobody perceives its output), then each cluster conform a universal code within itself. In the latter case, the entropy of the structure can increase if the agents within a cluster perceive the outputs of all other agents also within their clusters (periods where $H(\Xi)$ increases on in Fig. 9).

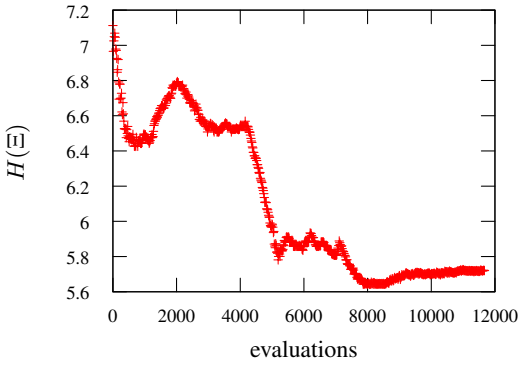


Figure 9: Evolution of the entropy of the population structure.

This is exemplified by the code distribution and population structure we obtained (see Fig. 10).

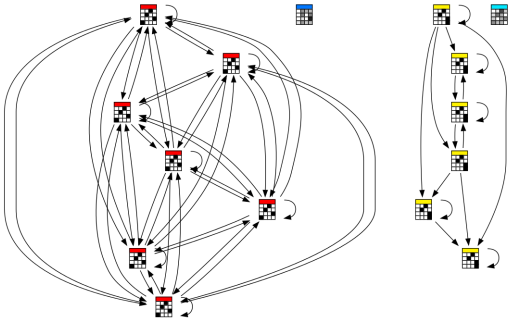

Figure 10: Each node in the graph corresponds to the code of an agent. There is a weighted edge between agent $\theta_i$ and $\theta_j$ if $p(\theta_i, \theta_j) > 0$ (which is the weight). The temperature colours on top of the nodes indicate the amount of environmental information they would contribute to any agent perceiving only that agents output.

## Discussion

We considered three different scenarios of code evolution: in the first one, all agents perceived the outputs of all other agents, including itself. We argued that two main stages of evolution can be recognised: in the first stage, a universal code is established, which can be optimal or not. If it is not optimal, then a second stage will achieve optimality. The same result was obtained in (Vetsigian et al., 2006), in a model of the evolution of the genetic code (represented as a probabilistic mapping between codons and amino acids), although universality and optimality were simultaneously achieved.

In the mentioned work, which developed further the ideas of (Woese, 2002, 2004), the authors argue that the universality of the genetic code is a consequence of early communal evolution, mediated by horizontal gene transfer (HGT) between primitive cells. In this evolutionary process, they argue, larger communities will have access (through the exchange of genetic material) to more innovations, leading to faster evolution than smaller ones. Then, "it is not better genetic codes that give an advantage but more common ones" (Vetsigian et al., 2006). Although their model does not explicitly show this property, it is captured in our model. We show that a more common, but not optimal code is widely adopted within a population (see Fig. 6). However, in our model, a code imposes itself as universal not because it provides access to more innovations (in our model there is no "code exchange", only the outputs are shared), but because the population structure forces the adoption of the most popular code. After this stage, further changes in the code of the agents eventually lead to optimality.

In another related work, (Oudeyer, 2005) explored the origins of language in a scenario consisting of artificial agents with a coupled perception and production of speech sounds. Although this work is focused on plausible mechanisms for the origin of language, it assumes the same similarity principle as we do (hearing a vocalisation increases the probability of producing similar vocalisations), arriving to the same outcome (a universal language, or code).

Our second scenario, where the structure of the population is a grid, showed how establishing local conventions in early stages of evolution constrains the outcome of the code distribution, since to reconcile different conventions, several simultaneous changes are needed. On the other hand, in our third scenario, where we let the structure of the population change simultaneously with the codes themselves, such situations are avoided by "disconnecting" clusters with dissimilar conventions. This property enhances evolution, and can potentially lead to the adoption of several different conventions within an (increasingly fragmenting, or "speciating") population.

The evolution and establishment of conventional codes as defined by Barbieri could be interpreted, in the widest sense, as a form of cultural evolution. While communication between individ-

uals of a population opens up the possibility of "signal cheaters", our model does not allow such behaviour, since the code producing the outputs functions, implicitly, as the interpreter of the perceived signals.

## Conclusion

In the proposed model, we introduced a key assumption which allowed us to evolve, for some structures, universal and optimal codes. This assumption states that an agent cannot distinguish the sources of the outputs it perceives from other agents. Following from this, a universal code will necessary introduce semantics by relating symbols to environmental conditions (via the internal states of the agent). Our model proposes an information-theoretic way of measuring the similarity within a population of codes.

In this work, we proposed, as an evolutionary principle, that agents try to maximise their side information about the environment indirectly by maximising their mutual code similarity. This behaviour produces several interesting outcomes in the code distribution of a structured population. Depending on the population structure, it captures the evolution of a universal and optimal code (well-mixed population structure), while also the evolution of different codes organised in clusters (in a freely evolving structure), which allows the establishment of optimal as well as suboptimal conventions.

## References


Barbieri, M. (2003). The organic codes-An introduction to semantic biology. *Genetics and Molecular Biology*.

Barbieri, M. (2008). Biosemiotics: a new understanding of life. *Die Naturwissenschaften*, 95(7):577–99.

Battail, G. (2009). Applying Semiotics and Information Theory to Biology: A Critical Comparison. *Biosemiotics*, 2(3):303–320.

Deely, J. (2006). On 'semiotics' as naming the doctrine of signs. *Semiotica*.

Donaldson-Matasci, M. C., Bergstrom, C. T., and Lachmann, M. (2010). The fitness value of information. *Oikos*, 119(2):219–230.

Favareau, D. (2007). The evolutionary history of biosemiotics. *Introduction to biosemiotics*, pages 1–67.

Görlich, D., Artmann, S., and Dittrich, P. (2011). Cells as semantic systems. *Biochimica et biophysica acta*, 1810(10):914–23.

Hansen, N. and Kern, S. (2004). Evaluating the CMA evolution strategy on multimodal test functions. *Parallel Problem Solving from Nature-PPSN VIII*.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–95.

Heil, M. and Karban, R. (2010). Explaining evolution of plant communication by airborne signals. *Trends in ecology & evolution*, 25(3):137–44.

Igel, C., Heidrich-meisner, V., and Glasmachers, T. (2008). Shark. *Journal of Machine Learning Research*, 9:993–996.

Kelly, J. (1956). A new interpretation of information rate. *IEEE Transactions on Information Theory*, 2(3):185–189.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Oudeyer, P. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*.

Platt, T. G. and Fuqua, C. (2010). What's in a name? The semantics of quorum sensing. *Trends in microbiology*, 18(9):383–7.

Seger, J. and Brockmann, H. J. (1987). What is bet-hedging? *Oxford surveys in evolutionary biology*.

Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423.

Vetsigian, K., Woese, C. R., and Goldenfeld, N. (2006). Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, 103(28):10696–10701.

Waters, C. M. and Bassler, B. L. (2005). Quorum sensing: cell-to-cell communication in bacteria. *Annual review of cell and developmental biology*, 21:319–46.

Woese, C. R. (2002). On the evolution of cells. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8742–7.

Woese, C. R. (2004). A new biology for a new century. *Microbiology and Molecular Biology Reviews*, 68(2):173–186.

Wong, a. K. and You, M. (1985). Entropy and distance of random graphs with application to structural pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 7(5):599–609.