Punctuated evolution and robustness in morphogenesis

D. Grigoriev

CNRS, Mathématiques, Université de Lille
Villeneuve d'Ascq, 59655, France
dmitry.grigoryev@math.univ-lille1.fr

J. Reinitz[1,2,3,4]

[1] *Department of Statistics,*
University of Chicago, Chicago, IL 60637, reinitz@galton.uchicago.edu
[2] *Department of Ecology and Evolution,*
University of Chicago, Chicago, IL 60637
[3] *Department of Molecular Genetics and Cell Biology,*
University of Chicago, Chicago, IL 60637
[4] *Institute for Genomics and Systems Biology,*
University of Chicago, Chicago, IL 60637

S. Vakulenko

Institute for Mechanical Engineering Problems, Bolshoy pr. V. O.61
Sanct Petersburg, and ITMO University, Sanct Peterburg Russia
vakulenfr@mail.ru

A. Weber

Computer Science Department, University of Bonn
Bonn, Germany
weber@cs.uni-bonn.de

Abstract

This paper presents an analytic approach to pattern stability and evolution problem in morphogenesis. This approach is based on the ideas of the gene and neural network theory. We assume that gene networks contain a number of small groups of genes (called hubs) controlling morphogenesis process. Hub genes represent an important element of gene network architecture and their existence is empirically confirmed. We show that hubs can stabilize morphogenetic pattern and accelerate the morphogenesis. The hub activity exhibits an abrupt change depending on the mutation frequency. When mutation frequency is small, these hubs suppress all mutations and gene product concentrations do not change, thus, the pattern is stable. When the environmental increases and the population needs new genotypes, the genetic drift and other effects increase the mutation frequency. For the frequencies larger than critical, the hubs turn off, and as a result, many mutations can affect phenotype. This effect can serve as an engine for evolution. We show that this engine is very effective: the evolution acceleration is an exponential function of gene redundancy. Finally, we show that the Eldredge-Gould concept of punctuated evolution results from the network architecture, which provides fast evolution, control of evolvability, and pattern robustness.

1. Introduction

Robustness is an important property of biological systems. Wild-type organisms are buffered, or “canalized”, against environmental or genetic variation in the course of both development (Manu et al., 2009b) and evolution (Rendel, 1959). This term was coined by C. H. Waddington (1942), who stated that ”developmental reactions, as they occur in organisms submitted to natural selection ... are adjusted so as to bring about one definite end-result regardless of minor variations in conditions during the course of the reaction”. More recent work has shed light on the mechanistic origins of canalization behavior. In some cases a specific gene, called a “genetic capacitor” is responsible for canalizing behavior (Bergman and Siegal, 2003; Levy and Siegal, 2008; Moczek, 2007) while in other cases the canalization behavior arises from a small network of genes (Manu et al., 2009a). These genes or networks of genes buffer environmental or genetic variations, thus canalizing pattern formation and evolution. This situation implies an apparent paradox, because canalized systems are nevertheless able to evolve successfully to adapt to environmental changes. Different mechanisms of canalization have been discussed, for example, in (Gunji and Ono, 2012; Gursky et al., 2012; Gunji et al., 2014). In the paper by Gunji and Ono (2012) a cellular automata-based model is proposed to generate a French flag pattern as a model of canalization due to agents equipped with sociality. In this model cell can be considered as an agent that transports morphogen. A pattern occurs as a result of interaction of neighboring agents. The paper by Gursky et al. (2012) contains a review of the canalization problem, describes different canalization mechanisms and considers canalization as a result of gene interactions. Here we develop these ideas suggesting that the basis of robustness and canalization is a network architecture in which the genes or small networks responsible for canalization are hubs in larger gene networks. Reminds that the hubs are strongly connected nodes in networks (Albert and Barabási, 2002), a key element of network architecture. In systems level studies of metabolic networks, this architecture has been described as "bow-tie" connectivity Zhao et al. (2006). Here we use a variant of this idea, called an "empire structure" by Vakulenko (2013), wherein highly connected hubs play the role of organizing centers, and each center interacts with many weakly connected nodes, called satellites (see Fig. 1). Note that hubs are a universal feature of scale-free networks, and have been identified in a wide variety of natural and human-generated networks, including the genetic, metabolic and economic networks, as well as the internet (Albert and Barabási, 2002; Lesne, 2006). In particular, centralized networks have been empirically identified in molecular biology, where the centers can be, for example, transcription factors, while the satellite regulators can be small regulatory molecules such as microRNAs (Li et al., 2010) or the target genes of the *Drosophila* segmentation network.

In this work we address the apparent paradox of evolutionary change in the face of canalizing stability by means of an analytically tractable mathematical model of the canalization effect and its abrogation. We show that populations which experience an increase in mutation rate or pass through a bottleneck, possibly because of environmental stress, release hidden genetic information. Specifically, hubs in the network can create an abrupt change. transition. In normal conditions, the hubs stabilize the gene expression pattern against all mutations. When the mutation rate increases or the population size N decreases, the mutation probability p becomes higher (in particular, this effect can result from the genetic drift effect, since the drift induces a noise of intensity proportional to $N^{-1/2}$). In particular, there is a threshold effect in which when the probability p of a mutation becomes more than some critical value p_c, i.e., $p > p_c$, the hub stabilization fails. This effect produces an abrupt change and it is sharper for large gene networks. At this point hidden genetic information is manifested in phenotype, so that hidden mutations captured during a long stability period start begin to play a role in creating new phenotypes better adapted to new ecological conditions. This evolutionary mechanism corresponds to the punctuated evolution ideas of Gould and Eldredge (1972). Here we show that punctuated evolution is a natural consequence of an empire-type network organization.

In our approach to this problem, we also address a serious difficulty in evolutionary theory, namely that classical population genetics does not consider the hierarchical organization of multicellular organisms into differentiated cell types that in turn make up tissues and organs. It instead treats the organism as a unitary entity possessing organismal fitness. This formulation avoids the problem that a condition that may be selectively advantageous for a particular cell type with a particular gene expression state is selectively disadvantageous for another cell type with a different gene expression state, where both cell types contain the same genetic material. In this work we represent the selective effects of mutations on a hierarchically organized multicellular organism by using an idea from theoretical computer science Valiant (2009). Namely, in this paper L.G. Valiant has found a connection between the evolution problems and the fundamental computer science problem $P \neq NP$ Cook (1971) that helps to formulate the problem in a more rigorous mathematical way.

Our approach to the problem of understanding punctuated equilibrium in the evolution of multicellular organisms in the face of canalization is based on an analogy between these evolution processes and hard-combinatorial problems. In the last decades these problems have received great attention from mathematicians and theoretical physicists (Friedgut and Bourgain, 1999; Deroulers and Monasson, 2006; Achlioptas, 2001; Mertens et al., 2006; Mézard and Zecchina, 2002). This analogy enables us to obtain an analytical relation for the evolution speed as a function of gene redundancy.

The connection of evolution with hard combinatorial problems allows us to formulate a precise statement of the meaning of *feasible* evolution (Valiant, 2009). In the framework of this model, evolution is feasible if one can find a local search algorithm, for example, a greedy one, that resolves the problem in P$oly(n)$ elementary steps. These steps can be thought of as single mutations. This idea, which connects the $P \neq NP$ problem with evolutionary biology was first formulated by Valiant (2009) (note that from this paper we have used only idea about connection between evolution feasibility and algorithm feasibility, and also a connection with problem $P \neq NP$, our model significantly differs from model Valiant (2009)).

Note that if P = NP, an equality that most do not believe to be correct, then evolution is always feasible.

The most fundamental hard-combinatorial problem is the famous k-SAT one. We state some important facts about k-SAT in the next section.

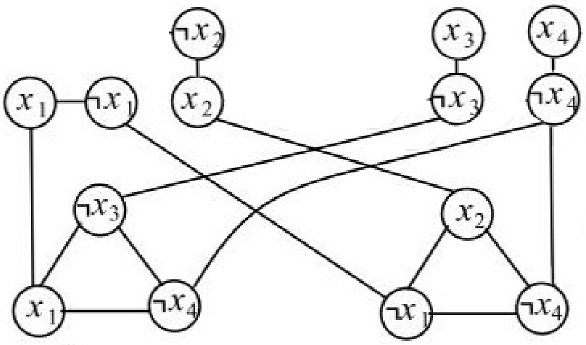
2. *k*- SAT problem

2.1. Formulation of the problem

Cook and Levin (1971) have shown that k-SAT problem is NP-complete, moreover, this problem has important applications for bioinformatics.

The k-SAT problem can be formulated as follows. Let us consider the set $V_n = \{x_1, ..., x_n\}$ of boolean variables $x_i \in \{0, 1\}$ and a set $\mathcal{C}_m$ of m clauses. The clauses C_j are disjunctions (logical ORs) involving k literals $y_{i_1}, y_{i_2}, ..., y_{i_k}$, where each y_i is either x_i or the negation $\bar{x}_i$ of x_i. The problem is to test whether one can satisfy all the clauses by an assignment of boolean variables.

It can be illustrated by the following picture (see Fig 1).



$$\phi = (x_1 \vee \neg x_3 \vee \neg x_4) \wedge (\neg x_1 \vee x_2 \vee \neg x_4)$$

Figure 1: This image illustrates a toy $k - SAT$ problem for $n = 4$ logical variables x_1, x_2, x_3, x_4 for $k = 3$ and $m = 2$. Clauses are triangles, we have here two clauses of length 3. Solution of the problem is correct, if all clauses are true. In this case we can take, for example, $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1$. This toy example is easy to resolve, however, the problem becomes difficult for $n, m >> 1$. In our biological model, we interprete x_i as genes and clauses as gene patternsfor different cell types.

A biological interpretation of k-SAT is quite transparent and can be formulated as follows. The number n is the gene number. Each gene is involved in formation of many differentiated cell types, and the gene can be either turned on or turned off in a given cell type. We have m cell types, therefore, the case $m \gg 1$ corresponds to the formation of a multicellular organism with many cell types. The number m can be interpreted, therefore, as a rough measure of the phenotype complexity. The parameter k determines genetic redundancy; that is, a cell type is formed by k different genes. The main difficulty of the k-SAT problem is that a logical variable x_i can be involved in different clauses as x_i and $\bar{x}_i$ therefore, it is difficult to assign x_i in a correct way. Biologically, this effect corresponds to the pleiotropy of genes. Namely, activation of a gene can help to create a useful cell type but, on the other hand, can become deleterious in other cell types.

Fig.1 show that, despite this model simplicity, cells can have different gene expression patterns, for example, if $x_1 = 1$, gene x_1 is not expressed for 1 cell type and expressed for 2-th one. It also illiustrates the simplest case of the gene pleiotropy effect. In fact, the activation of the first gene, $x_1 = 1$, helps in expression of 1-th cell type, but, at the same time, it prevents to create 2th one. We obtain the 2-th cell type due to the gene redundancy (on Fig 1. by a choice of x_2, x_3 and x_4).

2.2. Random large k-SAT models

The famous unresolved problem of theoretical computer science, P $\neq$ NP, is equivalent to the following question: does there exist an algorithm solving the k-SAT problem in polynomial time, that is to say in Poly($|X|$) steps, where X is the k-SAT problem input and $|X| = nm$ the size of this input. Suppose we are dealing with a randomly generated k-SAT formula and $n \gg 1$. The parameter $\alpha = m/n$ plays a key role in the description of k-SAT asymptotic behaviour as $n \to \infty$.

Evolution can be considered as a problem involving a number of constraints. A natural mapping of this problem to computer science is the problem of making an assignment in a random formula in Conjuctive Normal Form (CNF) to satisfy the maximal possible number of m disjunctions. If the disjuctions have the same size k, we obtain the classical k-SAT problem. Many properties of k-SAT also hold for more general constraint problems, and it is useful to briefly discuss random k-SAT.

Consider k-SAT of a random structure with $m = \alpha n$ clauses and n variables assuming $n \gg 1$. Let $k > 2$. For $\alpha > \alpha_c(n, k)$, where α_c is a critical value, there are no solutions on average with probability close to 1, and for $\alpha < \alpha_c(n, k)$, there exists, on average, a solution with probability close to 1. By "on average", we mean that we always consider a formula, that is to say a scheme of genetic regulation, of a random structure. A naturally plausible conjecture is that there exists a limit

$$\lim_{n\to+\infty} \alpha_c(n,k) = \alpha_c(k).$$

This result is not yet proved, but a close result has been demonstrated (Friedgut and Bourgain, 1999; Achlioptas, 2001):

$$\alpha_c(n,k) \approx \ln(2)2^k, \quad k >> 1, \quad n \to \infty.$$

There also exists an important second critical value $\alpha_d(n,k) < \alpha_c(k)$. For large k this value can be estimated by

$$\alpha_d(n,k) \approx 2^k/k.$$

If $\alpha < \alpha_d(n,k)$ then all solutions form a unique cluster. A cluster in the space $\{0,1\}^n$ of boolean sequences of length n can be defined as a connected set. Here we assume that two sequences are connected (adjacent), if the Hamming distance between them equals 1 (so, the sequences differ in a single coordinate).

If the solutions form a single big cluster, this means that one solution can be obtained from another by flipping some number of boolean variables. In this case simple algorithms of local search are capable of finding solutions in Poly(n) time. There are a number of such algorithms (WalkSat, GSat, DPLL etc.) (Selman et al., 1992; Kirkpatrick and Selman, 1994; Mézard and Zecchina, 2002). Local search algorithms will have difficulties beyond the clustering phase transition Mézard and Zecchina (2002). Thus, evolution proceeds while the level of the gene freedom is sufficiently high. Note that there is a very effective algorithm, Survey Propagation (SP) (Braunstein et al., 2005; Mertens et al., 2006), that allows us to find a solution very rapidly; for $n = 10^6$ it proceeds in minutes.

If $\alpha_c(n,k) > \alpha > \alpha_d(n,k)$, the set of solutions is a union of exponentially many clusters.

2.3. Evolution and random k-SAT

As stated above, we identify the clauses (centers) with cell types, logical variables with genes, and the parameter k with gene redundancy. Thus the problem of understanding the evolution of of cell types and tissues becomes a hard combinatorial problem. It is natural to assume that such problems have a random structure because evolution is a random process.

We conjecture that evolution may be successful, when the number of cell types m is less than $\alpha_d n$, where n is the number of genes. Then the solutions form a single cluster and hence all solutions can be obtained by a local search in $Poly(n)$ steps. This threshold α_d grows exponentially with k, implying that the phenotype complexity grows exponentially as a function of redundancy.

Note that we do not think that evolution can be effective in a domain where many clusters coexist. The transition from one cluster to another faces a fitness barrier in the sense that it must fail to satisfy hn clauses for some $h > 0$ (Moore and Mertens, 2011).

3. Model

3.1. Boolean gene network model

We consider a picture of gene regulation where the genes $u_1, \dots, u_n$ are Boolean variables, $u_i \in \{0, 1\}$, so that $u_i = 0$ means that gene i is turned off and $u_i = 1$ means that it is turned on (Kauffman, 1969; Valiant, 2009; Thieffry and Thomas, 1995). The total pattern of gene expression is denoted by $u = (u_1, ..., u_n)$. We assume that each cell type z_j of an organism is controlled by many genes. The formation of cell type j under the control of u is denoted by $z_j = 1$.

Clearly that the genetic regulation may be very complicated. In general, we are dealing here with a complicated boolean function $z_j = f_j(u_1, ..., u_n)$. For example, we can try to design f_j as a multilayered perceptron, since they are universal approximators and therefore can simulate all boolean networks. For example, for two layers we have

$$z_j(u) = \sigma(\sum_k W_{jk}\theta_k(u) - \tilde{h}_j), \quad \theta_k(u) = \sigma(w_{k1}u_1 + ... + w_{kn}u_n - h_k), \quad (3.1)$$

where $\sigma(z) = 1$ for $z > 0$ and $\sigma(z) = 0$ for $z \leq 0$ and $w_{ji} \in \{-1, 0, +1\}$.

In the simplest case, where we have only a single layer (perceptron model), we can set $W_{jk} = \delta_{jk}$, where δ_{jk} stands for the Kronecker symbol and $\tilde{h}_j = 0.5$. Then z_j is given by

$$z_j(u) = \theta_j(u) = \sigma(w_{j1}u_1 + ... + w_{jn}u_n - h_j). \quad (3.2)$$

We suppose that the threshold $h_j = \sum_{w_{ji}<0} w_{ji} + \xi_j$, where $\xi_j \in (0, 1)$ is a uniformly distributed random number. At the level of computer science, the step function σ and the threshold h_j ensures the disjunctive property of the expression (3.2). At the level of biology, we suppose that the values of w_{ji} denote the influence of the controlling genes u_i on terminal differentiation genes that are not explicitly included in the model, such as actin and myosin in a muscle cell, ion channels in a neuron, and so on. We assume that each w_{ji} takes the value 1 or –1 with equal probability $\beta/2n$, with $\beta > 0$ a parameter. Thus the quantities z_j are disjunctions of a random subset of literals $u'_1, ..., u'_n$. Each literal u'_k can be either a variable u_k or its negation $\bar{u}_k$. A negation occurs in z_j if a gene inhibits the formation of the j-th cell type. Note that each disjunction involves, on average, $k = \beta$ literals. The parameter β thus represents the level of redundancy. Note that we use the disjunctions because these logical functions have an important advantage: they provide a maximal redundancy effect. Note that if we choose h_j in another way, we can obtain other logical operations, for example, majority functions.

Let us consider an example. Consider 3 cell types and 10 genes. Then, by adjusting some w_{ij}, one can obtain, for example, such relations:

$$z_1 = u_3 \; OR \; \bar{u}_5 \; OR \; u_7, \quad z_2 = u_4 \; OR \; u_5, \quad z_3 = \bar{u}_1 \; OR \; u_8 \; OR \; \bar{u}_4 \; OR \; u_9. \tag{3.3}$$

The negation $\bar{u}_k$ is appeared because the corresponding $w_{jk} < 0$, i.e., k-th gene inhibites j-th type of cell differentiation. This example shows a gene pleiotropy effect.

Note that this simple model fails to explain robustness. We improve it in section 4 where a model describing canalization is proposed. This model involves two layers.

3.2. Evolution

We represent fitness *in silico* by the sum of extant cell types needed for survival, which we write in terms of the model as a sum of satisfied clauses

$$W_F(u) = \sum_{j=1}^{m} w_j z_j(u), \tag{3.4}$$

where z_j are defined by (3.2) and $w_j > 0$ are weights such that $w_1 + w_2 + ... + w_m = m$. We set, for simplicity, $w_j = 1$.

Evolution requires variance in the population on which natural selection operates. We generate variation in the model by a process of simple Boolean mutations.

We consider a population u consisting of N_{pop} members. Each member of the population u^l is represented by the logical variable $u^l = \{u_1^l, u_2^l, ..., u_n^l\}$. An elementary mutation step is to flip one bit of $u_{j(l)}^l$ in the lth population member for all $l = 1, ..., N_{\text{pop}}$, where we choose j_l randomly from the indices $\{1, ..., n\}$, such that all choices are independent. If the mutation increases the fitness $W_F(u^l)$ we keep the new value u^l, otherwise we retain the old one.

We allow selection to operate from time to time as follows. The only organisms conserved in the population after a selection step is the individual $u^* = u^l$ with maximal fitness W_F and all mutants that can be obtained from u^* by mutations in k_{loc} places. After selection we again continue the mutation process as above. This algorithm is a simple variant of the well known GSAT (Selman et al., 1992; Kirkpatrick and Selman, 1994).

After a large number of mutations the system loses the property of evolvability. That is, it reaches a state u_{final} in which fitness can no longer increase because mutations for this final pattern are either neutral, conserving fitness so that $W_F(u_{\text{mut}}) = W_F(u_{\text{final}})$, or lead to a decrease in fitness with $W_F(u_{\text{mut}}) < W_F(u_{\text{final}})$ (see section on numerical simulations). This pattern fragility occurs for any simple algorithms of local search.

The question of how to buffer maximally adapted organism against deleterious mutations is nontrivial. We address this question in the following subsections.

4. Canalization by hubs

To describe canalization mechanism, we extend the simplest k-SAT model. To obtain such a more sophisticated model, we take into account the following basic fact: the hubs can stabilize the morphogenesis and buffer mutations (Bergman and Siegal, 2003). A way to describe this effect is to consider two layered perceptron (3.1) involving an extra term $Z_c(u)$ which describes the effect of hub genes involved in canalization:

$$Z_c(u) = \sigma(v_1 u_1 + ... + v_n u_n + h_0), \tag{4.1}$$

where v_i are some appropriate coefficients. Here we assume that, in (4.1), either $v_i = 1$ (activation of i-th gene by the center), or $v_i = -1$ (inhibition of i-th gene by the center). A mutation, that flips v_i, occurs with a probability p. This probability, together with the threshold h_0, will be defined by a specific procedure to be described below. Our choice of this buffering term is natural from a genetic point of view, where it corresponds to buffering via hubs, an effect confirmed experimentally (Bergman and Siegal, 2003; Levy and Siegal, 2008)). It is also a natural choice in terms of computer science because it can be obtained from the Valiant-Vazirani-Vazirani isolation lemma Valiant and Vazirani (1986). Such a term describes an action of a "central" gene acting on all genes that determine morphogenesis (see Fig. 1).

The difference between the simplest k -SAT local model and this two layer model can be explained by ideas Gunji and Ono (2012), in particular, by an interesting economical analogy suggested in this paper. In the k-Sat model we have only market agents interacting locally. Acting in a local manner, they form a global market (pattern), which may be unstable. In the extended model we take into account some global centers (for example, a government) which influence all market agents.

Then two layer boolean model with extra term Z_c is as follows:

$$z_j(u) = \sigma((1-b)\theta_j(u) + bZ_c(u) - \tilde{h}_j), \quad \tilde{h}_j \in (0,1), \tag{4.2}$$

where θ_j is defined by (3.2) and $b \in (0, 1/2)$ is a buffering parameter. In a sense, it is a minimal model that can produce the canalization and an abrupt change.

Finally, we have the following biological picture. Each cell type is controlled by a block of random structure consisting, on average, of β genes. Moreover, an organizing center, or "hub" influences the morphogenesis process via the term bZ_c involving a number of genes. Here the coefficient b determines the level of buffering, so that for larger b buffering is bigger. We call this an "empire structure" because the hub exerts control over the patterning process in the same manner as the center of an empire controls the periphery.

We now discuss the choice of v_i. To describe an abrupt change in evolution, the term Z_c should satisfy the following property:

Z: *Let $u = u^*$ be a maximally fit genetic pattern (without mutations). Then, if the mutation frequency is less than a critical level p_c, i.e. $p < p_c$,*

$$Z_c(u) = 1 \quad for \ all \ u \neq u^*.$$

If $p > p_c$,

$$Z_c(u) = 0 \quad for \ all \ u \neq u^* \ and \ Z_c(u^*) = 1.$$

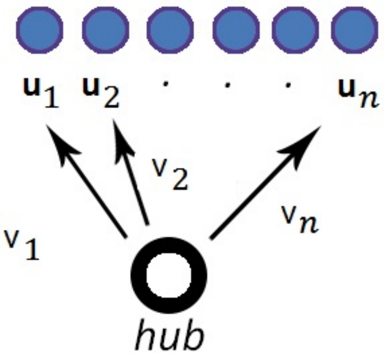


Figure 2: This image illustrates an hub action on genes and an "Empire structure". Genes u_i are "satellites", which are under control a center that influence satellites with intensities v_i.

The canalized fitness, that corresponds to (4.2, is then as follows:

$$W_{Fc}(u) = W_F(u) + bZ_c(u), \tag{4.3}$$

where $W_F(u)$ is the fitness for the k-SAT model.

4.1. The Hebb rule and canalization

We construct v_i in the buffering term $Z_c(u)$ as follows. Hebb's rule is a well known idea in biological learning theory. It states that if two neurons tend to fire synchronously, synaptic connections between them tend strengthen. Proposals have been made that such a process operates in evolution (Watson et al., 2010; Adams, 1998). This idea is particularly reasonable if the pair of genes considered as a code for transcription factors. These proteins bind to noncoding control DNA, and it is well known that such regulatory regions evolve rapidly with the frequent turnover of binding sites He et al. (2011). If there is a selective advantage for a pair of such genes to be simultaneously expressed, mutations that stabilize such interactions will provide the type of buffering effect we represent here as a generalization of Hebb's rule.

If we assume that, for the interaction between the genes and the center, an analogue of the Hebb rule holds, and, moreover, that the center is always active, we obtain the following algorithm:

A1: *If* $u_i = 1$ *over a long time period, then* $v_i = 1$*; otherwise* $v_i = -1$.

Evolution takes place over a long time, and so the algorithm can be recast into a statement about the final, maximally adapted state:

A2: *If for the final pattern* u *we have* $u_i = 1$*, then* $v_i = 1$*; otherwise* $v_i = -1$.

Algorithm **A2** leads to the relation

$$Z_c(u, X) = \sigma_a(h_0 - S), \quad S = \sum_{i=1}^{n} |u_i - u_i^*|, \tag{4.4}$$

where $\sigma_a(x)$ is a sigmoidal function, for example, $\sigma_a = (1 + \exp(-ax))^{-1}$ and u^* is the maximally adapted gene expression pattern. The quantity S is the Hamming distance between the current gene expression pattern u and the maximally adapted pattern u^*. Notice that S and thus Z_c involve all genes. Eq. (4.4) can be interpreted as follows: in the gene network there is a hub, which summarizes the contributions of all genes to stabilize pattern. The terms $w_i = |u_i - u_i^*|$ in (4.4) states a form of Hebb's rule. If all $u_i = X_i$, then $w_i = 0$, otherwise $w_i > const > 0$. If all $w_i = 0$ we have $Z_c = 0$. If h_0 is small enough in this case $Z_c = 1$.

In relation (4.4) we assume that there is a single hub. Such situation is unstable with respect to mutations, which can delete the hub. One can generalize (4.4) for the case of many (say, r) hubs that makes the hub system stabler with respect to hub deletions. Then

$$Z_c(u, X) = \sum_{l=1}^{r} \sigma_a(h_{0l} - S_l), \quad S_l = \sum_{i=1}^{n} \eta_{il} |u_i - u_i^*|, \tag{4.5}$$

where η_{il} are non-negative weight coefficients. In this study we consider the case $r = 1$.

5. Canalization and passage through bottleneck

We now show that, under certain assumptions, that there is a critical mutation frequency p_c

$$Z_c(u, p) \approx 1, \quad (p < p_c), \quad Z_c(u, p) \approx 0, \quad (p > p_c). \tag{5.1}$$

This means that as p increases gradually, the canalization term $Z_c(u, p)$ decreases sharply.

To this end, we compute the expectation $E[S]$ and the deviation $D[S]$ assuming that all u_i can mutate independently with a small probability p. Then $E[|u_i - u_i^*|] = p$ so $E[S] = np$. Also, deviation $E[|u_i - u_i|] = p(1-p)$ and hence $D[S] = np(1-p) \approx np$. For large n one has $(D[S])^{1/2} << E[S]$. The central limit theorem shows then that $h_0 - S$ is a gaussian random variable sharply localized at $h_0 - np$, thus, (5.1) holds. For the critical probability p_c we obtain

$$p_c \approx \frac{h_0}{n}, \quad (n \gg 1, \ np \ll 1). \tag{5.2}$$

We find, therefore, a mechanism, which is consistent with experimental data and fundamental biological concepts (Levy and Siegal, 2008; Bergman and Siegal, 2003; Manu et al., 2009b; Moczek, 2007). This can be understood in terms of a passage of the population through a bottleneck, perhaps induced by a stress.

Experimental data show. (Levy and Siegal, 2008; Bergman and Siegal, 2003) that a stress may damp the canalization effect and this, in turn, can lead to a loss of robustness. This fact can be explained by this model as follows. When a population consisting of N members, lives in a stress conditions, the population abundance N diminishes. The genetic drift effect is proportional to $N^{-1/2}$ (Sviregev and P, 1982). Therefore, for small N, when the population goes through a “bottleneck” (hard environmental conditions diminishing the population size), the mutation frequency p_{mut} exceeds p_c. This releases a hidden gene information captured at long stage when the population have lived in stable environment that, in turn, can produce formation of new organisms (more adapted to new hard conditions). Certainly, the stress can directly affect the probability p, if it is connected with chemical reagents, radiation etc.

6. Results of numerical simulations on evolution

6.1. Simulation of evolution and mutation effects

We have considered populations with $50 - -100$ members. Redundancy parameter β was taken in the interval $\beta \in [6, 10]$. During evolution process, the number m of constraints in a random CNF formula increasing from $m = 300$ up to $m \in (800, 1800)$. This means that we simulate an increasing, in time, sharpness of ecological environment and increasing organism complexity: an "organism" should satisfy more and more constraints. Between these steps of environment changes, we have made $100 - -200$ mutations for such population member. After each mutation series we have checked the fitness values, the pattern stability level and value of the buffering term.

Our results are as follows. First, they are consistent with main facts about k-SAT described in the previous section. We have found an abrupt change for $\alpha \in (10, 16)$, i.e., in the most of cases, a gene system with $\beta = 7$ and $n = 100$ genes can satisfy $m \in (1000, 1600)$ constraints (to create an organism with ≈ 1500 features). Solutions form a giant cluster and the Hamming distance between different solutions is not large (of order 10).

Notice that for $k = 7$ the best estimate for algorithmic phase transition is ≈ 33, see Achlioptas (2001). It is clear that in the our case this value should be smaller, since we have many clauses (disjunctions) of size < 7, and the algorithm is not optimal.

Patterns u^*, that maximize the fitness $W_F(u)$, are fragile if the buffer mechanism is turned out ($b_c = 0$). To check the pattern robustness we have produced 50 single random mutations and 50 random double mutations. We have observed the following picture. Let $\Delta W_1 = W(u^*) - W(u_{mut})$ be a vector of lengths 50 that contains the fitness variations for single mutations, and ΔW_2, ΔW_3 analogous vectors for double and triple mutations. Typical results for the last stage of evolution, when the constraint number $m = 1900$, are as follows. The quantity ΔW_3 lies within the interval $[3, 11]$, ΔW_2 lies within $[1, 9]$ and ΔW_1 is in $[0, 3]$, where we have $\approx 5 - 10$ neutral mutations (with $\Delta W_1 = 0$). It is natural that single mutations are less dangerous a part of these mutations are neutral. This negative mutation effect can be completely compensated by the buffering term $b_c Z_c$, which takes the values 50 for all mutations. For this final evolution stage, the averaged number of the satisfied constraints in the population (i.e., the fitness) is $W_{\text{averaged}} = 1887$ and the maximal number is $W_{\max} = 1892$. For previous evolution stages, when $m = 800$, we have $\Delta W_1 \in [-1, 3]$.

We observe here more neutral mutations and even some positive mutations. Namely, we have 3 positive mutations $\Delta W_1 = -1$ and ≈ 18 neutral mutations. The maximal number of the satisfied constraints is close to the limit value, $W_{max} = 799$.

Figs. 6.1 and 6.1 illustrate some properties of evolution.

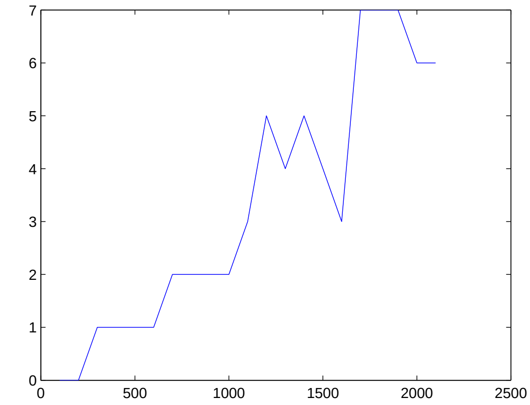


Figure 3: The plot of minimal (over the population) number of non-satisfied constraints m_{NS} as a function of the number of constraints m. Horizontal axis is m, $m \in [300-2000]$, vertical axis is the m_{NS} value. Each evolution step adds 100 new constraints. The number of mutation at each step 100, the population contains 100 member. The parameter $\beta = 7$

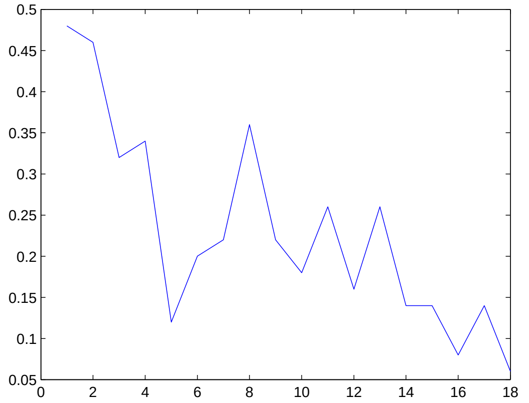


Figure 4: This plot shows a part of neutral double mutations as a function of the constraint number m ("organism complexity"). The parameter $\beta = 8$. On the horizontal axis values $m/100$ are shown.

6.2. Evolution boundaries

For usual k-SAT when all clauses have the same size K we can obtain a simple estimate of the critical value α_c of the phase boundary by the Markov estimate (Moore and Mertens, 2011). Let Z be the number of satisfying assignments for k-SAT. Then the Markov inequality says that $Pr\{Z > 0\} \leq EZ$. Computing the expectation we find that

$$Pr\{Z > 0\} < \xi^n, \quad \xi = 2(1 - 2^{-k})^{\alpha}.$$

The expected number of satisfying assignments is exponentially small if $\xi < 1$ that gives the for the phase boundary α_c the estimate $\alpha_c < 2^k \ln 2$.

We are dealing with a CNF formula of a random structure where the clauses have sizes $k_1, ..., k_n$ distributed by a normal law. Repeating the above arguments we obtain that the expected number of satisfying assignments is exponentially small if

$$S(\alpha, n) = \sum_{i=1}^{m} \ln(1 - 2^{-\mathrm{k_i}}) < -n, \quad m = \alpha n. \tag{6.1}$$

The equation $S(\alpha, n) = -n$ gives a rough estimate for the phase boundary α_c.

6.3. Evolution rate

It is clear that the evolution rate decreases in m and increases in β. One can compute the averaged number N_{uns} of non-satisfied constraints when we substitute a random assignment in k-SAT. This gives $N_{uns} \approx \sum_{i=1}^{m} 2^{-k_i}$, where k_i are clause sizes. The relation is consistent with numerical simulations. One can assume that for $b_c = 0$ the rate R_{ev} of our evolution algorithm(R_{ev} is the number of mutations steps in order to satisfy all constraints) is proportional to N_{uns}. One can suppose that the following rough approximation holds:

$$R_{ev} \approx \text{const} \sum_{i=1}^{m} 2^{-k_i} \approx \text{const } m \ 2^{-\beta}. \tag{6.2}$$

The coefficient const was adjusted by numerical simulations and the least square method. The precision of this approximation is about $25 - -30\,\%$, see Figs. 4 and 5.

7. Conclusion

In this paper, we have presented an analytic approach to canalisation and evolution problems. This approach is based on ideas of theoretical computer

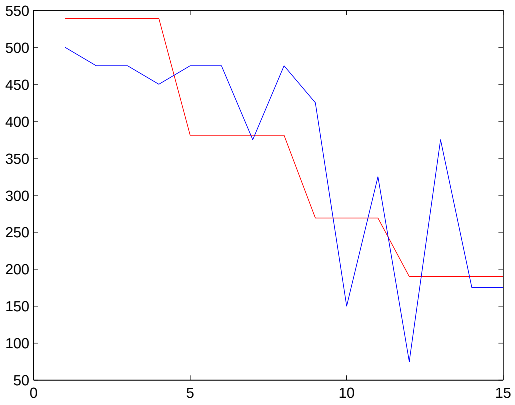


Figure 5: This plot shows the dependence of evolution rate r on the number of constraints m. The blue curve presents results of numerical simulations, the red curve is const $\cdot\, 2^{-\beta}$, where the constant is adjusted by the least square method. The vertical axis shows the number of steps to satisfy all the clauses. Here $n = 100, \beta = 6$ and $m \in [300, 600]$. The blue curve exhibits a high randomness in the step number.

science, gene and neural network theory. We assume that gene networks contain some hubs, which control canalization processes (this assumption is confirmed by experimental data). The hub activity depends on the mutation frequency. For a favorable environment, when the mutation frequency is small, these hubs block all mutations and gene product concentrations do not change, thus, the pattern (phenotype) is stable. When environmental conditions become harder and the mutation frequency increases, the hub control fails. The natural mechanism for this increases is genetic drift, which grows as the population abundance diminishes. Moreover, the environment can increase the mutation frequency directly, for example, by a radiation, chemical reagents etc. When the mutation frequencies become larger than a critical value, the hub control turns off. Then all hidden mutations can be released affecting the phenotypes. This effect serves as an engine for evolution.

The second key question is to explain whether this engine is powerful enough to form the complicated patterns during a “reasonable” (i.e. sufficiently short) time, since the time for evolution is limited. If a species fails to create new phenotypes during a ”short” time period this species will

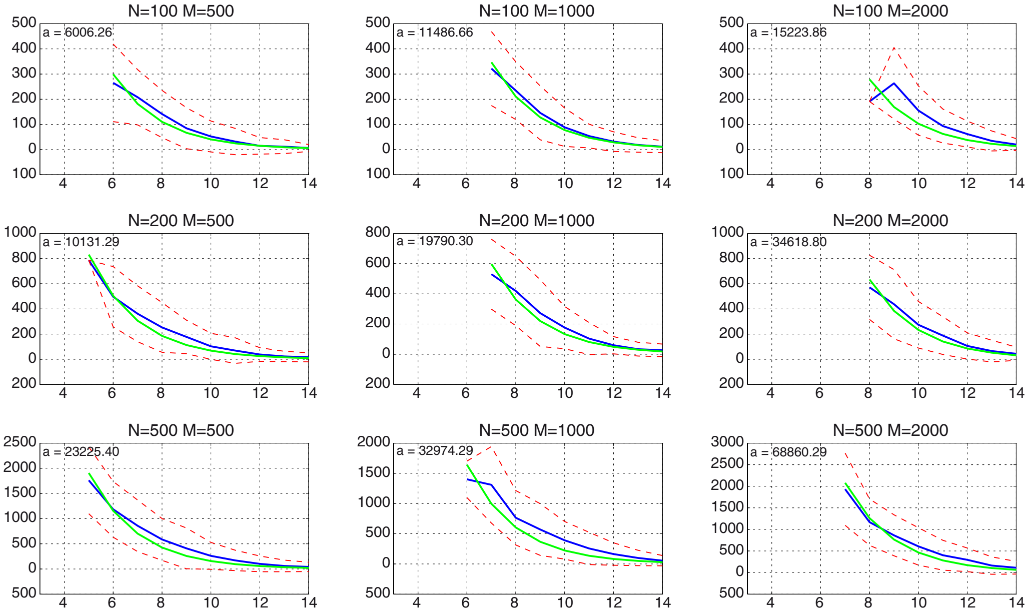


Figure 6: This plot shows the fitting of the evolution rate for different $m = M$ and $n = N$. We present the rate (the step number) as $a \exp^{-\beta/2}$, where a is fitted by the least square method. The horizontal curve shows the values of β. The green curve shows $a\ \exp^{-\beta/2}$, the blue one shows the mean values $\bar{s}$ of the numerical simulations s. The two dotted red curves are the plots of $s^{\pm} = \bar{s} \pm \sigma(s)$, where σ denotes the standard deviation.

become extinct. To formalize this problem we formulate a mathematically correct definition of a ”short time”. For this purpose, we use an approach inspired by Valiant (2009). We consider the morphogenesis as a hard combinatorial problem. This problem involves two key parameters, β and α. The first one determines genetic network redundancy and the second one can be interpreted as a “gene freedom”. The evolution rate is *exponentially fast as a function of the redundancy parameter*. We have presented an approximation for this rate. Simulations and theoretical results show that the evolution rate R_{ev} grows in genetic redundancy β in an exponential manner, as $R_{ev} \approx a \exp^{-\beta/2}$, where a is a constant. We conclude that a random evolution process is quite feasible when this genetic freedom is large enough, but when the number of constraints become too large, the evolution stops.

There appears an interesting connection of our model to the Gould and Eldredge theory of punctuated equilibrium. According to this theory, evolutionary change occurs relatively rapidly, alternating with longer periods of relative evolutionary stability Eldredge and Gould (1972). These ideas have excited the biocommunity—see the discussion in Ridley (2004) and recent paper Gunji et al. (2014), where, in particular, a fundamental connection of Gould and Eldredge theory with important concepts of adaptation and adaptivity is studied. According to Gunji et al. (2014), " adaptation reveals the passive mode of biological processes. In contrast, adaptability reveals the active mode of biological processes, the ability to thrive in an alternative environment that is hidden within living systems".

Our results show that gene networks with large gene redundancy and having a typical network architecture with hubs—which is found practically in all networks—can perform canalization and very fast evolution. Organisms controlled by such networks evolve in a punctuated manner. It is shown that strongly connected nodes (hubs) in gene networks are responsible for adaptivity wheres weakly connected genes perform adaptation.

Acknowledgements. The third author was supported by the grant of Russian Ministry of Education, 2012-1.2.1-12-000-1013-016 and his work was done in the frame of the project We 1945/7-1 funded by *Deutsche Forschungsgemeinschaft*. The third author was also financially supported by Government of Russian Federation, Grant 074-U01. The first author is grateful to Max-Planck Institut für Mathematik, Bonn for its hospitality. J.Reinitz was supported by US NIH grant RO1 OD010936 (formerly RO1 RR07801). The authors are grateful to B. Krüger for his help in performing some of the simulations.

Achlioptas, D., 2001. Lower bounds for random 3-SAT via differential equations. Theor. Comput. Sci. 265, 159–185.

Adams, P., 1998. Hebb and Darwin. J. Theor. Biol. 195 (4), 419–438.

Albert, R., Barabási, A. L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47– 97.
URL `http://rmp.aps.org/abstract/RMP/v74/i1/p47_1`

Bergman, A., Siegal, M. L., 2003. Evolutionary capacitance as a general feature of complex gene networks. Nature 424, 549–552.

Braunstein, A., Mézard, M., Zecchina, R., 2005. Survey propagation: An algorithm for satisfiability. Random Structures and Algorithms 27, 201–226.

Cook, S. A., 1971. The complexity of theorem-proving procedures. In: Proceedings of the third annual ACM symposium on Theory of computing. ACM, pp. 151–158.

Deroulers, C., Monasson, R., 2006. Criticality and universality in the unit-propagation search rule. Eur. Phys. Journal B 49, 339–369.

Eldredge, N., Gould, S. J., 1972. Punctuated equilibria: an alternative to phyletic gradualism. Models in paleobiology 82, 115.

Friedgut, E., Bourgain, J., 1999. Sharp thresholds of graph properties, and the *k*-sat problem. J. Am. Math. Soc.
URL `http://www.ams.org/jams/1999-12-04/S0894-0347-99-00305-7/`

Gunji, Y.-P., Ono, R., 2012. Sociality of an agent during morphogenetic canalization: Asynchronous updating with potential resonance. Biosystems 109, 420–429.

Gunji, Y.-P., Sakiyama, T., Murakami, H., 2014. Punctuated equilibrium based on a locally ambiguous niche. Biosystems.

Gursky, V., Surkova, S. Y., Samsonova, M., 2012. Mechanisms of developmental robustness. BioSystems 109, 329–335.

He, B. Z., Holloway, A. K., Maerkl, S. J., Kreitman, M., 2011. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet. 7 (4), e1002053.

Kauffman, S. A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol. 22 (3), 437–467.

Kirkpatrick, S., Selman, B., 1994. Critical behavior in the satisfiability of random boolean expressions. Science 264, 1297–1301.

Lesne, A., 2006. Complex networks: from graph theory to biology. Lett. Math. Phys. 78, 235–262.

Levy, S., Siegal, M., 2008. Network hubs buffer environmental variation in Saccharomyces cerevisiae. PLoS Biol. 6 (11), pe264.

Li, L., Xu, J., Yang, D., Tan, X., Wang, H., 2010. Computational approaches for microRNA studies: a review. Mamm. Genome 21 (1-2), 1–12.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., Reinitz, J., 2009a. Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. PLoS Computational Biology 5, e1000303.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., Reinitz, J., 2009b. Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. PLoS Biology 7, e1000049.

Mertens, S., Mézard, M., Zecchina, R., 2006. Threshold values of random k-sat from the cavity method. Random Struct. Algorithms 28 (3), 340–373.

Mézard, M., Zecchina, R., 2002. Random k-satisfiability problem: From an analytic solution to an efficient algorithm. Phys. Review E E 66, 056126–26.

Moczek, A., 2007. Developmental capacitance, genetic accommodation, and adaptive evolution. Evol. Dev. 9 (3), 299–305.

Moore, C., Mertens, S., 2011. The Nature of Computation. Oxford University Press.

Rendel, J. M., June 1959. The canalization of the *scute* phenotype of *drosophila*. Evolution 13 (4), 425–439.

Ridley, M., 2004. Evolution. Blackwell Publishing.

Selman, B., Levesque, H. J., Mitchell, D. G., et al., 1992. A new method for solving hard satisfiability problems. In: AAAI. Vol. 92. pp. 440–446.

Sviregev, J. M., P, P. V., 1982. Foundations of Theoretical Genetics. Nauka, Moscow.

Thieffry, D., Thomas, R., 1995. Dynamical behaviour of biological regulatory networksII. Immunity control in bacteriophage lambda. Bull. Math. Biol. 57 (2), 277–297.
URL `http://link.springer.com/article/10.1007/BF02460619`

Vakulenko, S., 2013. Complexity and evolution of dissipative systems. de Gruyter, Berlin.

Valiant, L. G., 2009. Evolvability. J. ACM 56 (1), 1–21.

Valiant, L. G., Vazirani, V. V., 1986. NP is as easy as detecting unique solutions. Theor. Comput. Sci., 2–7.

Waddington, C. H., 1942. Canalization of development and the inheritance of acquired characters. Nature 150, 563–565.

Watson, R., Buckley, C. L., Mills, R., Davies, A., 2010. Associative memory in gene regulation networks. In: Proc. of the Alife XII Conf. MIT Press, Odense, Denmark.

Zhao, J., Yu, H., Luo, J., Cao, Z., Li, Y., 2006. Hierarchical modularity of nested bow-ties in metabolic networks. BMC Bioinformatics 7, 386.