Structure of a Well-Known Modularity-Inducing Problem Domain*

Author One
Institution
Omitted, Omitted
abc@def

Author Two
Institution
Omitted, Omitted
def@def

Author Three
Institution
Omitted, Omitted
ghi@def

ABSTRACT

This is where the abstract goes.[1]

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:
Author One, Author Two, and Author Three. 2018. Structure of a Well-Known Modularity-Inducing Problem Domain. In *Proceedings of the Genetic and Evolutionary Computation Conference 2018 (GECCO '18).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Why modularity is important

Why the Wagner-Espinosa Soto model is important, and why we are also considering Larson's variant. We will almost certainly want to cite [2].

Why we need to understand the fitness landscapes

Sensitivity to algorithm variants

2 BACKGROUND

This is where we give more detail of the background in modularity.

We might also put in a brief description of our initial experiments with tournament selection, explaining that we were puzzled why we didn't see any modular solutions this way.

In the inception phase of this project, we utilised the Louvain heuristics to compute the partition of the network vertices in order to maximize the modularity of the given graph [**?**]. We applied the tournament selection scheme with the tournament size being three and the elitism mechanism with ten elites in every generation. As a result of this setting, the partition of the gene regulatory networks by the Louvain heuristics demonstrated a very low modularity score. As Figure X indicates, by simulating the work in [2], we had expected there would be a spike after 500 generations on modularity. In contrast, we observed a modularity decrease as a result.

Figure X. An example of evolutions that did not evolve out high modularity

In order to understand this puzzling phenomenon, we removed the elitism mechanism and changed the tournament to proportional selection scheme. In consequence, we eliminated the deviant phenomenon as Figure X indicates. Therefore, we hypothesized that the elitism mechanism or the tournament selection scheme hamper the evolutionary process on evolving out modular structures.

3 METHODS

This is where we describe what we're going to do. As discussed previously, we don't mention symmetry or noisy evaluation in this paper, nor do we cover hotspots, diploidy or dominance.

What we do work with is a basic GA with mutation as per previous work, and cross over. The variations are:

Espinosa Soto vs Larson fitness function Fitness proportionate (roulette) selection vs tournament (maybe a couple of different tournament sizes)

We may end up merging this with the experiments section

We utilize genetic algorithms as our evolutionary simulation tools. The gene regulatory network that we used in this paper was originally proposed by Wagner [8] and customized by Espinosa-Soto and Wagner [2] as well as Larson et al. [5].

All simulation code was implemented in Java 1.8.0 and Python 2.7.10. They are all publicly available at https://github.com/xxxxxxxxxx. Modularity was evaluated using the NetworkX package with the community API [3]. All the generated data can be downloaded at: https://drive.google.com/file/xxxxxxxxxxxxxxxxxx.

3.1 Model

Cells in an organism display heterogeneity in functionalities and morphologies, while they contain the same set of genes. In other words, cells interpret the same genetic material in different ways so that their behaviours and structures vary. These distinct interpretations are due to the regulation via the activation and repression of genes [8]. In brief, effects of different genes are not mutually independent. A protein that is generated by a gene may activate or repress other genes. A gene regulatory network can be a mathematical directed graph to express these relationships of genes in an organism [8]. Specifically, genes can have two different patterns, namely activation and repression. The term "gene activity pattern" is adopted to represent the activeness status of the entire set of genes. Different gene activity patterns mean the distinct cellular functions and forms [2].

*Produces the permission block, and copyright information
[1]This is an abstract footnote

We re-constructed the model that was utilized in the work done by Espinosa-Soto and Wagner, which is a model to represent a gene regulatory network [2]. In this model, a gene regulatory network with N genes will be in the form of an adjacency matrix $A = a_{ji}$, which acts as a genotype of an individual. Each entry a_{ji} is restricted to be either 1, 0 or -1, which represents an activation, absence or repression interaction from gene j to gene i, respectively. The gene activity pattern of this network at time t can be expressed as a Boolean row vector $s_t = [s_t^0, ..., s_t^{N-1}]$. A certain gene i can either be active ($s_t^i = 1$) or inactive ($s_t^i = -1$). The transition of state activity is modelled by the equation below

$$s_{t+\tau} = \sigma[\sum_{j=1}^{N} a_{ji} s_t^j] \tag{1}$$

where $\sigma(x)$ equals 1 if $x > 0$ and is 0 otherwise.

3.2 Fitness

The fitness here evaluates the likelihood that an attractor is obtained when facing perturbations [2]. In other words, Espinosa-Soto and Wagner imposed a bias of robustness on their gene regulatory network models in order to indirectly select modular networks. This is because modular networks can limit perturbations in a module so that the overall structure will not be heavily affected [1]. That is, more modular networks are more robust.

There are two or more stages in their experiments on discovering the conditions under which modularity starts emerging. In the first stage, gene regulatory networks are evolved under selective pressure towards regulating a particular gene activity pattern, while facing some perturbations. The original gene activity pattern before perturbation is called a target. In the second and further stages, networks are evolved under selective pressure to regulate new gene activity patterns, while preserving the ability to regulate the old patterns. In the particular case where there were two gene activity patterns, the first stage lasted for 500 generations and the second took another 1500 generations.

The perturbations of targets are randomly generated in every generation when evaluating the fitness of gene regulatory networks. In Espinosa-Soto and Wagner's experiments, a network would face 500 perturbations comprising different corrupted versions of gene activity patterns. Each gene will have a probability of 0.15 to be perturbed into its opposite activity. A further study was conducted to explore a sufficient number of perturbations in order to shorten the computational time while maintaining a similar eventual improved modularity. It was concluded that 75 or 100 perturbations would lead to the noteworthy emergence of modularity [7]. Therefore, 75 perturbations are undertaken for evaluating the fitness of each gene regulatory network in order to reduce the running time.

Larson et al. applied another approach for evaluating the fitness of networks [5]. They generated a static set of perturbations at the beginning and utilised this same set of corrupted targets whenever network fitness was calculated. This method converts the original stochastic fitness evaluation into a deterministic one. That is, the evolutionary landscape of individuals under this fitness evaluation will remain unchanged in each generation. On contrast, Espinosa-Soto and Wagner's fitness evaluation will lead to the evolutionary landscape to shift every generation.

The fitness value of a gene regulatory network reflects its robustness in recovering from various perturbations. The error function compares an attractor of the network dynamics to the original gene activity pattern. That is, a successful network is able to regulate a corrupted pattern to its initial form. Then, the Hamming Distance G between the attractor and the original pattern was calculated. Previous experiments indicated that it normally took fewer than 20 transitions to reach the attractor [8]. Thus, non-stable attractors are assumed to be those gene regulatory networks that take more than 20 steps to attain the stability, or are cyclically stable. They are treated to have a maximum Hamming distance D_{max}. This is followed by a calculation of the contribution from each perturbation attractor to the fitness, which is defined as a developmental trajectory $\gamma = (1 - D/D_{max})^5$ [2]. Afterwards, this process is repeated to determine 75 γ_i, $1 \leq i \leq 75$. Finally, the fitness of a network is calculated as

$$f(g) = 1 - e^{-3g} \tag{2}$$

where g represents the artimetic mean of the sum of all γ_i [2]. As to cases where there are more than one gene activity patterns, the arithmetic mean of $f(g)$ for all the patterns was take. Consequently, a gene regulatory network with a high fitness is able to lead to different attractors matching different targets.

3.3 Evolutionary Simulations

Espinosa-Soto and Wagner imposed a bias towards low-density gene regulatory networks in mutation [2]. A node in the network has a probability $\mu = 0.05$ to mutate every generation, and it either can lose or gain an interaction. The probability for a node to lose an interaction can be calculated as

$$p(u) = \frac{4r_u}{4r_u + N - r_u} \tag{3}$$

where N is the number of gene nodes in a gene regulatory network, and r_u equals to the number of regulators of gene u [2]. That is, the number of genes that exert effects on gene u. In contrast, the probability for a gene u to obtain an interaction is defined to be $1 - p(u)$. That is, it can keep the sparseness of the network, which computational biology research suggests is necessary for the emergence of modularity.

Espinosa-Soto and Wagner did not apply a crossover mechanism in their simulation [2]. In the reconstructed model by Larson et al., they limited crossover to nine possible partition locations of a 10-node network, corresponding to nine possible rows for splitting the adjacency matrix of a network horizontally [5]. When two matrices A_1 and A_2 are selected for crossover at index i, matrices of their children will be produced as

$$C_1[0 : i-1, :] = A_1[0 : i-1, :]$$
$$C_1[i : 9, :] = A_2[i : 9, :]$$
$$C_2[0 : i-1, :] = A_2[0 : i-1, :]$$
$$C_2[i : 9, :] = A_1[i : 9, :]$$

However, this horizotal crossover may not only make the parental networks exchange modular clusters, but also exchange some interactions between the two modules. This may corrupt modularity. In contrast, we use a crossover mechanism that swaps interactions between modules in a gene regulatory network with connections

between modules in another network. Compared with the crossover mechanism of Larson et al., this approach, as Figure X illustrates, will better preserve the community structure (Wilcoxon signed-rank test; $p < 0.0372$).

3.4 Modularity Metric

We adopted the Q scoring system to quantify modularity in a network based on the algorithm proposed by Newman [6]. Briefly, this approach is defined as the difference between the ratio of the number of edges in the network connecting nodes within a module over the number of all the edges, and the same quantity when assigning the nodes into the same modules yet edges are assumed to be randomly connected in the network [4]. Formally, Q is calculated as

$$Q = \sum_{i}^{K} [\frac{l_i}{L} - (\frac{d_i}{2L})^2] \quad (4)$$

where i represents one of the K potential modules within a network, L is the total number of connections in a network, l_i stands for the number of interactions in the module i, and d_i is the sum of degrees of all the nodes in module i [2]. In other words, Q considers the two ratios of both intra-module connection density and inter-module connection density [6]. A network that is considered to be good on modularity must consist of as many within-module edges and as few inter-module edges as possible. However, it will result in $Q = 0$ if all the nodes are partitioned into the same module.

The value Q will sit in the range of $[-\frac{1}{2}, 1)$. Nodes in the gene regulatory network are partitioned into different groups according to their regulating gene activity patterns.

4 EXPERIMENTS

This may end up merged with the methods section. Detailed settings for the experiments, including full evolutionary tableaux.

5 RESULTS

This is where we present the detailed results. We need fitness and modularity results. We also need the comparison between the optimum and the fittests high-modularity solutions. Then we need the results of deleting non-modular links from optimal solutions, and comparing resulting fitnesses, and showing the fitnesses of intervening paths. Finally, it may be desirable to check the result of evaluating the fitness of a good solution under one sampling method from one generation with the fitness of that solution under some other fitness function (i.e. how much do the fitnesses of an individual vary from generation to generation under Espinosa-Soto evaluation? How different are the Larson fitnesses? Are the differences larger or smaller for modular or non-modular solutions?

6 EXPERIMENTS

This may end up merged with the methods section. Detailed settings for the experiments, including full evolutionary tableaux.

7 DISCUSSION

This is where we discuss the results, and their implications. Basically, we express this as puzzlement: we need to understand the fitness landscape better, and this probably requires better tools. We also discuss the sensitivity of the fitness landscape and the emergence of modularity, and compare it with biological evolution, where the emergence of modularity is robust and almost universal.

8 CONCLUSIONS

Summarise the results

Why these results are important.

Where we go from here.

ACKNOWLEDGEMENTS

?Do we need to acknowledge grants here? Assistance from Bongard? etc.

REFERENCES

[1] Alan Aderem. 2005. Systems biology: its practice and challenges. *Cell* 121, 4 (2005), 511–513.
[2] Carlos Espinosa-Soto and Andreas Wagner. 2010. Specialization can drive the evolution of modularity. *PLoS computational biology* 6, 3 (2010), e1000719.
[3] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX.* Technical Report. Los Alamos National Laboratory (LANL).
[4] Nadav Kashtan and Uri Alon. 2005. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America* 102, 39 (2005), 13773–13778.
[5] Ari Larson, Anton Bernatskiy, Collin Cappelle, Ken Livingston, Nicholas Livingston, John Long, Jodi Schwarz, Marc Smith, and Josh Bongard. 2016. Recombination Hotspots Promote the Evolvability of Modular Systems. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion.* ACM, 115–116.
[6] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113.
[7] Mariko Totten. 2015. Exploring the Evolution of Modularity in Gene Regulatory Networks. (2015). https://scholarworks.uvm.edu/cgi/viewcontent.cgi?referer=https://www.google.com.au/&httpsredir=1&article=1081&context=hcoltheses
[8] Andreas Wagner. 1996. Does evolutionary plasticity evolve? *Evolution* 50, 3 (1996), 1008–1023.