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

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

---

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

Unpublished working draft. Not for distribution.

2018-01-15 23:18. Page 1 of 1–4.

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/xxxxxxxxxxxxxxxxx.

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].

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, $1 \le i \le 75$. 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

$$Q = \sum_i^K [\frac{l_i}{L} - (\frac{d_i}{2L})^2]$$ (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

Gene activity patterns and the essential parameters of our evolutionary simulations are provided in the form of Tables 3.1 and 3.2 in order to facilitate repeatability of these experiments. The detailed explanations of these parameters are given after Table 3.2. Overall, only the elite number will be specified in each experiment, since only that may vary in different experiments. All the other parameters are specified in Table 3.1 and Table 3.2 and are consistent in the experiments.

Table 1: Table to test captions and labels

Gene Activity Pattern	Generation to Add a New Pattern
1, -1, 1, -1, 1, -1, 1, -1, 1, -1	0
1, -1, 1, -1, 1, 1, -1, 1, -1, 1	500

4.1 Diagonal Crossover Mechanism Promotes Modularity

I simulated 40 independent evolutions for the development with no crossover and with each of the two crossover mechanisms, namely horizontal crossover and diagonal crossover, respectively. None of these simulations applied elitism. Overall, the diagonal crossover mechanism performed better than no crossover and the horizontal crossover, regarding both regulatory performance and modularity emergence, as Tables 3.1 and 3.2 indicate.

The Boolean model that I have utilised to simulate biological networks was originally proposed by Wagner in his study on "epigenetic stability" [8]. His work indicated that random recombination made no difference for the evolution of stability, which may be due to the freeness of random recombination on choosing locations to undertake crossover. This can corrupt the modular structures in biological networks.

Conversely, my experimental results suggested that proper recombination methods can contribute to the evolvability of organisms. The diagonal crossover proposed in this report is able to preserve underlying network modules. Although the crossover mechanism utilised by Larson et al. did not preserve community structures as well as diagonal crossover, its partitioning is still based on a network-like structure. This can be the reason why both of these two crossover mechanisms could help in obtaining modularity, with diagonal crossover better than horizontal crossover. Meanwhile, different combinations of parental traits can increase the diversity of the population so that the evolution can be more exploratory.

4.2 Greed Hampers Modularity

4.2.1 Proportional Exceeds Tournament Selection on Generating Modularity.
We simulated
4.2.2 Elitism Hampers Modularity.
We simulated 40 evolutionary trials with 10 elites and without any elites. That was 80 trials in total. The experimental results indicate that elitism will hamper both the networks' regulatory capabilities and modularity emergence, as shown in Table 3.3 and Table 3.4.

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 ANALYSIS

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

7 DISCUSSION

7.1 Modular systems did not gain dominance on survivability

Greedy methodologies, including elitism and tournament selection scheme, impede the emergence of modularity under our evolutionary simulations. This implies that individuals who performed optimally in the early stage might not be optimal on modularity. In

other words, the most competitive elites in each generation did not have the most modular gene regulatory networks.

Overall, these phenomena suggest that the modularity emergence condition, namely gene specialization promotes modular networks, may not be plausible to explain biological modularity. They indicated that modules in the simulated gene regulatory networks did not gain dominance in determining the survivability of individuals. However, biologically, modular networks are dominant and ubiquitous. In order to further investigate the plausibility of this theory, namely specialization driving modularity, we obtained the most optimal gene regulatory network among networks that were the most modular. Conversely, we also collected the network that was the least modular among those that had the greatest fitness value. These networks were collected from the generated results of experiments in Section X.X.

Biologically, I expected the fitness value of the latter would be lower than the fitness of the former. Nevertheless, the situation was converse. That is, some less modular networks were more robust than more modular ones, as Table X.X indicates. This is not consistent with what has been observed in biology.

Initially, I hypothesized that the inconsistency was due to the targeted gene activity patterns being over-simple. That is, the number of genes in a pattern was not sufficient or the number of patterns was not enough. A modular network may give great performance on complex tasks, but worse than non-modular ones for simple tasks. Thus, I conducted a complicated evolutionary simulation consisting 7 patterns, each of which comprised 15 gene nodes. This evolution lasted for 35,000 generations. Other detailed parameters are in the Appendix X.

The evolutionary progress for the best-fit individual in every generation is in Figure 4.1. I conducted the modularity dominance analysis again and the results are in Table 4.4. Overall, the complex of gene activity patterns could not resolve the issue of non-dominance for modular networks on survivability.

7.2 Inter-Module Connections Can Hamper Network Fitness

Fitness values of gene regulatory networks were measured after removing interconnections between modules in order to understand the functionality of inter-module interactions. The results indicated that among 40 networks which had the highest fitness values and relatively low modularity Q scores in their corresponding evolutionary simulations, 24 of them demonstrated higher fitness after manually converting them into modular structures by deleting inter-module edges. That is, there existed non-modular networks that exhibited better fitness performance after removing all the inter-module connections. For example, the right network in Figure X.X was the consequence of removing inter-module connections of the network in the left. The fitness value of the latter was 0.9502 after it had removed 6% connections of the former, whose fitness was 0.9472. Further statistical investigations will be conducted in the future.

Originally, we suspected that this deviance was due to the fact that these modified solutions had a lower density than was expected from the evolutionary operations (sec 2.4), and thus may have been excluded from the search space. Nevertheless, further investigation

revealed that the average number of edges for those networks that increased fitness values after triming their inter-module connections was approximately 30. That is, it was not due to the bias on the sparseness that caused this anomaly.

In order to further comprehend this phenomenon on why our evolutionary simulations could not find a path to the trimed networks, we recorded the fitness value of removing one inter-module edge in turn, until deleting all of them. We plotted graphs as Figure X.X, where x-axis represents the number of inter-module edges that have been discarded, y-axis represents the corresponding fitness values. Interestingly, most of our collected plots demonstrated a steady increasing trend for fitness vs deleting edge numbers, whereas genetic algorithms could not find these paths.

7.3 Fluctuant landscapes are essential for generating modularity

The stochastic fitness evaluation used by Espinosa-Soto and Wagner [2] demonstrated much higher fitness and modularity Q score than Larson et al.'s deterministic fitness evaluation. Therefore, we hypothesized that a fluctuannt landscapes for individuals during the evolution might be necessary to develop high modularity. In order to verify this hypothesis, we collected

7.4 More modular networks require fewer connections

8 CONCLUSIONS

Summarise the results
Why these results are important.
Where we go from here.

ACKNOWLEDGEMENTS

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.