UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Evolution of Modularity in Biological Signalling Networks

Mestrado em Bioquímica
Especialização em Bioquímica

Daniel Vilar Jorge

Dissertação orientada por:
Professor Francisco Pinto

2015

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Evolution of Modularity in Biological Signalling Networks

Mestrado em Bioquímica
Especialização em Bioquímica

Daniel Vilar Jorge

Dissertação orientada por:
Professor Francisco Pinto

2015

Agradecimentos

Em primeiro lugar quero, agradecer à minha Família, nomeadamente os pais, irmã e avós, por terem sido eles que, para além de me financiaram todos os estudos até agora, me suportaram em todas as alturas difíceis deste percurso académico e sempre acreditaram nas minhas capacidades.

Quero também dar destaques a algumas pessoas que, não fazendo parte diretamente da minha família, sempre estiveram presentes na minha vida como tal, nomeadamente a família Santos Guia e em especial ao meu mais antigo amigo, Diogo Guia pela motivação e paciência para me ouvir nas alturas mais críticas. Também agradeço a outro amigo de infância, João Paulo Bento, por todo o apoio profissional que me deu, por todas as *jam sessions* e conhecimentos variados que me foram importantes toda a vida.

Quero ainda agradecer ao Samuel Lourenço Jacob por me sempre me fazer soltar uma gargalhada nos momentos em que seria impensável tal acontecer, à Sofia Vargas, por me incentivar a continuar com o trabalho quando a vontade não era muita, ao Rui Paiva por compor músicas espetaculares que me acompanharam durante a escrita da dissertação e me dar na cabeça por não acabar o meu trabalho mais depressa, ao Viriato Queiroga por acreditar no meu trabalho, fazer-me ter vontade de ir atrás dos meus objetivos e obrigar-me, de uma maneira cordialmente agressiva, a terminar a dissertação.

Por último, quero agradecer ao Francisco Pinto por toda ajuda que me deu, todo o conhecimento que me passou, por toda a paciência que demonstrou para comigo durante o período em que trabalhámos juntos e por acreditar nas minhas capacidades.

Abstract

The concept of modularity associated to signalling and gene expression regulatory networks has been a field of study for the past 20 years. There are a few theories that try to explain the origin of this architecture in biological networks but there is no agreement on which factors contribute more and are actually responsible for modularity to arise as a predominant topology. One of the ideas that is most accepted in the scientific community is that if a population of organisms is exposed between every few generations to new environmental challenges, as long as they have common sub-problems, that should be enough for modularity to be selected as the preferred network architecture type.

In this work, we try to approach this problem in a different fashion, where our digital organisms population is represented by a set of directed graphs and they were exposed to different environments during their lifetime. Our objective is to prove that this condition alone is enough for modularity to arise on the population's signalling networks. Also we evaluate the influences of mutational parameters and their individual contributions, as well as the impact in terms of number of environments and how similar they are on the evolution towards modular networks.

Our results show that it is possible to have an evolution towards modular networks just by exposing organisms to different environmental challenges through their life time, and not necessarily with environments with common sub-problems. Additionally, faster gene duplication rates and a slower gene interactions mutation rates are important in this process. Gene elimination does not seem to have any impact. This work also shows that fitness and modularity are not directly correlated, even if modularity may represent an evolutionary advantage, their evolution patterns can be different. Notably, the same simulation conditions that makes possible for modularity to arise, can also produce high fit populations that are not modular.

Resumo

A modularidade é um conceito abstrato que é usado para descrever um tipo de arquitetura em grafos genéricos que pode ser aplicado a redes de sinalização e regulação da expressão genética. Um grafo é modular quando é possível dividir os seus vértices em grupos (*clusters*) com um grande número de ligações entre os vértices que constituem esse *cluster* e um baixo número de ligações entre vértices que pertencem a *clusters* diferentes. Em redes biológicas é geralmente associado à ideia de que proteínas ou genes com funções relacionadas têm muitas ligações entre si, por exemplo, mecanismos regulatórios comuns. Pelo contrário, proteínas com funções distintas tendem a ter poucas ligações entre si.

Este conceito associado a redes de sinalização e regulação da expressão genética tem sido uma área de estudo durante os últimos 20 anos, e existem várias teorias que tentam explicar os motivos que estão na base da origem deste tipo de arquitetura em redes biológicas. Umas defendem que a modularidade interfere diretamente com a capacidade de sobrevivência do organismo e assumem que desta forma, a seleção natural seja o fator decisivo. Outros assumem que não existe este tipo de pressão seletiva e que, por exemplo, a duplicação de nós é o principal fator, enquanto outros defendem que é o ambiente que faz com que os organismos desenvolvam redes modulares. No entanto, não existe um consenso de quais são os fatores que são efetivamente responsáveis pelo aparecimento desta organização. Uma das propostas mais aceites na comunidade científica é a de que uma população de organismos, entre gerações, sendo exposta a novos desafios ambientais, desde que estes tenham objetivos e estímulos parcialmente semelhantes, acaba por selecionar a modularidade como o tipo de arquitetura de redes preferencial.

É assumida, neste trabalho, a hipótese de que a modularidade é consequência de variações ambientais. No entanto, o problema é abordado de uma maneira diferente. Neste trabalho simulamos populações de organismos em que cada indivíduo, no seu tempo de vida, é exposto a múltiplos ambientes.
De modo a alcançar os objetivos, foram feitas várias simulações de evolução com populações de organismos artificiais durante 10000 gerações. O tamanho de população foi fixado para 1000 indivíduos por geração e sua reprodução foi sincronizada. Cada indivíduo é representado por um grafo dirigido, que modela uma rede booleana de sinalização e regulação da expressão genética, ou seja, cada vértice só tem dois estados: 0 (inativo) e 1 (ativo). Estes vértices foram classificados como vértices de estímulo, vértices intermédios e vértices de resposta.

Os vértices de estímulo mimetizam proteínas que têm um papel sensorial para com o ambiente, como recetores de membrana. Estes só podem ter ligações saída para vértices intermédios. É impossível para os nós de estímulo terem qualquer tipo de ligação com vértices de resposta, ou outros vértices de estímulo. Cada organismo tem um número fixo de dezoito vértices de estímulo que não podem ser duplicados nem eliminados por nenhum tipo de evento mutacional. Cada ambiente é definido por uma lista de nós de estímulo que são ativados num indivíduo quando é exposto a este e o estado dos nós de estímulo é imutável num dado ambiente.

Os vértices intermédios mimetizam proteínas que regulam a atividade de outras proteínas ou interferem, na expressão genética, tais como cinases em vias de sinalização ou fatores de transcrição. Estes só podem ter ligação de entrada a partir de outros vértices intermédios e dos vértices de estímulo e podem ter ligações de saída que visam outros vértices intermédios ou vértices de resposta. É impossível formarem ligações de entrada a partir de vértices de estímulo e ligações de saída para vértices de estímulo. O número deste tipo de vértices pode variar a cada geração por eventos mutacionais de duplicação e eliminação de vértices e também podem diferir entre organismos da mesma geração, excetuando a primeira, em que o número inicial para todos os organismos é definido para um vértice intermédio desligado de todos os outros vértices da rede.

Os vértices de resposta mimetizam proteínas com uma variedade de funções celulares, tais como enzimas, proteínas estruturais ou transportadores. Estes só podem formar ligações de entrada a partir de vértices intermédios e não podem ter quaisquer ligações de saída. É impossível um vértice de resposta estar ligado a um vértice de estímulo. Cada organismo tem um número fixo de dezoito vértices de resposta que, tal como os vértices de estímulo, não podem ser duplicados nem eliminados por nenhum evento mutacional. Para cada ambiente é definida uma lista de nós de resposta que devem estar ativos para que o organismo tenha um nível de adaptação ótimo a esse mesmo ambiente.

Toda a população é, então, exposta aos diferentes ambiente para determinar os que melhor se adaptam e em seguida é feito um sorteio dos indivíduos que figurarão na geração seguinte, sendo que os mais adaptados têm maior probabilidade de aparecer em maior proporção. São, ainda, aplicados processos de duplicação de vértices, adição ou remoção (mutação) das interações entre vértices e eliminação de vértices. O número de ambientes a que os organismos foram expostos, dependendo do ensaio, foram de 2, 4 e 8, sendo que ainda podiam ter sobreposição de estímulo e objetivos, ou não.

O objetivo é o de provar que a exposição dos organismos a diferentes ambientes durante o seu tempo de vida é suficiente para que se observe modularidade nas redes de sinalização da população. Os resultados recolhidos suportam esta ideia. Para além disso, foram feitos estudos, variando individualmente cada um dos parâmetros mutacionais, mostrando que estes têm impacto quer da adaptabilidade de um indivíduo, quer na modularidade de sua rede de sinalização, bem como na quantidade de vértices e interações que apresentavam. Uma taxa de duplicação mais rápida será importante para adquirir maior adaptabilidade e maior modularidade no caso de organismos enfrentarem um problema ambiental mais

complexo (mais sobreposição entre ambientes e maior número dos mesmos), uma vez que serão necessários mais vértices para o resolver. A taxa de eliminação de vértices da rede não apresentou ter algum efeito. Já a taxa de mutação de interações, observou-se que, sendo mais lenta, pode contribuir para uma maior adaptabilidade aos ambiente e também à arquitetura modular das redes; nunca poderá ser demasiado rápida de modo a que as redes consigam evoluir sem perder todas as suas características entre gerações.

Quanto ao impacto da mudança de ambientes na evolução das redes de sinalização, observa-se que quanto maior é o numero de ambientes e mais sobrepostos eles estão, mais gerações leva até que os indivíduos se comecem a adaptar ao meio e a ter, de forma mais clara, uma rede mais modular. Também é possível observar que nestes cenários ambientais complexos, existe um desacoplamento entre a evolução da adaptabilidade e modularidade, já que evoluem de forma e em alturas diferentes.

Uma descoberta interessante é a de que as redes de indivíduos extremamente adaptados não apresentam, obrigatoriamente, uma arquitetura modular. É ainda possível chegar a redes de sinalização com arquiteturas diferentes da modular, e altamente adaptados aos ambientes, em que os parâmetros mutacionais utlizados foram idênticos, bem como o tipo de exposição ambiental. Isto é uma prova de que a modularidade não está diretamente relacionada com a adaptabilidade de um indivíduo ao meio, mas este tipo de arquitetura, pode, efetivamente, contribuir para um melhor perfil de adaptação.

Todos os modelos e processos biológicos foram computados utilizando a linguagem de programação Python 2.7 e a análise estatística foi feito na linguagem R.

Conteúdo

Lista de Figuras

Lista de Tabelas

List of Equations

Capítulo 1

Introduction

1.1 Signalling and gene expression regulatory networks

Living organisms have to adapt to changing environmental conditions [1], such as nutrient availability or temperature variations, for example. Because of that, they had to develop signalling and genetic expression regulation networks, that translate the environmental stimuli to physiological responses, such as varying the concentration of certain proteins or change their activity [2].

These networks can be represented as graphs that are constituted by both nodes, that can be proteins or genes, and the regulatory connections that are established between them [3, 4, 5].

One goal of systems biology is to find the design principles of these networks [4, 5, 6]. These design principles are usually associated with some evolutionary advantage and functional adaptation [7].

In biological networks, one of the design principle that has been described is modularity [8], but the evolutionary advantage and mechanisms that led to modular networks are not fully described [9].

1.2 Modularity

1.2.1 What is modularity?

Modularity is an abstract concept used to describe a type of architectures in generic graphs that can be applied to signalling or gene expression regulatory network. A graph is modular when is possible to divide it's nodes in groups (clusters) with a high number of connections within the nodes that constitute a cluster and a low number of connections between nodes that belong to different clusters. In biological networks this is often associated with the idea that proteins or genes that have related functions have a lot of

interactions between them, e.g. regulatory mechanisms, compared to proteins or genes with different functions [9, 8].

1.2.2 Origins of modularity

The origin of the modular architecture is a very controversial topic [8]. There have been several studies trying to explain how did organisms evolved modularity in signalling and regulatory networks, but there is not an agreement. However, it is probably caused by a number of forces acting in different contexts. In that optic it is necessary to identify both the forces and their contributions [9, 10].

Some models that try to explain the origin of modularity have natural selection driving it [8]. It has been proposed that modularity contributes directly to higher fitness, but these models lack a detailed analysis through the evolutionary process. This does not allows us to know what factors actually lead to the modular architecture [8].

Others say that, due to environmental pressures, there are some traits that need to change together frequently, integrating a module [8, 11, 10]. One of the most accepted hypothesis is that rapidly changing environments might be one of the reasons to arise this kind of architecture, as long as they share a common sub-problem [11, 10]. It has been shown [10] that varying environments every 20 generations can affect an organism's structure, robustness, genotype-phenotype mapping and speeding up the evolution.

Another possible explanation is called 'differential erosion of pleiotropic effects'. This means that there is a selection for robustness to noise that will eventually lead to modular networks, where the pleiotropic nodes lose that propriety [8, 12]. On the other hand, Tran and Kwon [13] say that network robustness is negatively correlated to modularity. This would be because if a gene that is part of a module suffers a perturbation, the other nodes that belong to that same module are more likely to suffer from this perturbation as well.

There are also models that do not account for natural selection as the main driving force of modularity. One of them is duplication-differentiation model [5] where the nodes of the network are proteins and the edges are interactions between these proteins. The idea is that the network grows by selecting a random node and duplicating it. When a node is duplicated, it inherits the edges that were connected to the original node but there is a probability to lose those connections and a probability to connect to new nodes. Only new replicated nodes can form or lose connections, and this process will, eventually, lead to modular networks [8].

In this work we follow the hypothesis that modularity is a consequence of environmental changes. In contrast with previous studies, where changes in the environment take place between generations [10, 11], we will study the impact of environmental changes within a generation. Every generation will be exposed to the same set of varying environments, excluding from our study the effects described in previous works.

1.3 Objectives

In this work, the main objective is to test how an organism reacts to a multi-environment exposition during its lifetime in terms of modularity, i.e. if this condition alone is enough to promote an evolution towards a modular signalling-network without the need of changing the environments between generations. Besides that, we wanted to study the influence of the:

- Number of environments
- Similarity degree between different environments
- Mutational parameters

on the signalling networks evolution.

Capítulo 2

Methods

2.1 Individual Organism and Population Definition

To achieve the objectives described in **Section 1.3**, we ran several evolutionary simulations of artificial organisms populations for 10000 generations. The population's size was fixed to 1000 individuals per generation and their reproduction was synchronized.

Each individual is defined by a directed network that models a signalling and gene expression regulatory boolean networks, i.e. all its node have only 2 states: **0** (inactive) and **1** (active). The nodes are classified in input Nodes ($\boldsymbol{I_n}$), intermediate Nodes ($\boldsymbol{N_n}$) and output Nodes ($\boldsymbol{O_n}$).

The input nodes mimic proteins that have an environment sensory role, e.g. membrane receptors. They can only have outward edges connecting to intermediate nodes. It is impossible for the input nodes have any kind of edge connecting them to an output node or other input nodes. Every organism have a fixed number of 18 input nodes that can not be duplicated or deleted by any mutational events. Each environment is defined by the list of input nodes (**Table 2.1 and 2.2**) that are activated when an individual is exposed to it and the input nodes' state is immutable in a given environment.

The intermediate nodes mimic proteins that regulate the activity of other proteins or interfere in the genetic expression, e.g. kinases in signalization pathway or transcription factors. They can only have inward edges from the input nodes and other intermediate nodes and can have outwards to output nodes and other intermediate nodes. It is impossible to form inward edges from output nodes or outwards to input nodes. Their number can vary each generation by duplication and deletion events and can be different in each organism within the same generation, apart from the first. In the starting population, the number of intermediate nodes in every organism is set to one and that node is not connected to any other nodes.

The output nodes mimic proteins with a variety of cellular functions, e.g. enzymes, structural proteins or transporters. They can only have inward edges from intermediate nodes and can not have any kind of outward edges to any other node. Every organism have

a fixed number of 18 output nodes that, like the input nodes, can not be duplicated or deleted by any mutational events. For each environment is set a list of output nodes (**Table 2.1 and 2.2**) that should be active on an organism that would have an optimal level of adaptation.

2.2 Fitness Estimation

In each generation, for every individual and every environment is made an environmental signalling propagation (input nodes activation by environmental factors like temperature or oxygen availability), until we get the organisms' response (output nodes activation). To simplify the model, every edge has a positive signal, i.e. corresponds to an activation. An output node is activated if it is the final node of a directed path within the network, starting on an activated input node. Every intermediate node that belong to directed paths that start in activated input nodes become active as well.

A list of the activated output nodes, for a given environment, can be compared to the expected ideal organism's response to that environment, allowing us to estimate a fitness measure using the Jaccard coefficient (**Equation 2.1**) [16]. The global fitness of an environment set, which is calculated using de product of the different fitness values of the environments that belong to that same set, is a quantification of the adaptive success of an individual that will affect the average number of descendants of that same individual that will figure in the next generation. To be able to compare the fitness values of individuals exposed to a different number os environments, it was necessary to normalize the values. We considered that no matter how many environments the organisms were exposed to, their fitness is evaluated 8 times. If an individual is exposed to 8 different environments (the maximum number of different environments in this work), each partial fitness corresponds to a different environment. If an individual is exposed just to 2 environments, then each one of them is evaluated 4 times. An example of this fitness estimation is presented in **Section 2.6**.

So, the **Fitness** equation will be:

$$\boldsymbol{Fit} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{J(F,E)}_{\boldsymbol{i}}^{\boldsymbol{T/n}} \tag{2.1}$$

where $\boldsymbol{J(F,E)}$ is the Jaccard coefficient [16] between an individual's phenotype, $\boldsymbol{F}$, and the environment's ideal phenotype, $\boldsymbol{E}$ (**Equation 2.2**). $\boldsymbol{n}$ is the number of the total environments of the set, on this work's simulations n={2,4,8} and T the number of environments to be normalized to, i.e. max(n). In this case, T = 8.

$$\boldsymbol{J(F,E)} = \frac{|\boldsymbol{F} \bigcap \boldsymbol{E}|}{|\boldsymbol{F} \bigcup \boldsymbol{E}|} \tag{2.2}$$

Arithmetically, for an individual to have a fitness different than zero, it has to have at least one stimulus on a phenotype node for each environment of the set. An example of how to estimate fitness can be found in.

2.3 Population Reprodution

The individuals that will figure in a next generation are chose by a sampling with replacement where each individual can be selected with a probability proportional to it's global fitness. These new individuals' networks can be affected by mutational events: mutation of the connections (edges) between nodes, node duplication and node elimination. Theses events allow a diversification of the population and the possibility of new network architectures.

2.4 Network Architecture

To evaluate the networks' architecture evolution, we account, for each individual: the number of nodes, the number of edges and the modularity, calculated in terms of nodes and edges.

Modularity base on nodes ($\boldsymbol{Mod_{node}}$) is define by the ratio between the number of intermediate nodes that are activated in a single environment ($\boldsymbol{a_{nodes}}$) and the number of nodes that were activated in, at least, one environment of the set ($\boldsymbol{A_{nodes}}$). The referred nodes on the numerator of the fraction are part of the ones on the denominator, so the ratio will vary between 0 and 1.

$$\boldsymbol{Mod_{node}} = \frac{\boldsymbol{a_{nodes}}}{\boldsymbol{A_{nodes}}} \tag{2.3}$$

To define the modularity base on edges ($\boldsymbol{Mod_{edge}}$), it is important to define an active edge as a connection between two activated nodes. The modularity based on edges is then defined as the ratio between the number edges that were activates in a single environment ($\boldsymbol{a_{edges}}$) and the edges that were activated in, at least, one environment of the set ($\boldsymbol{A_{edges}}$). An example of how to calculate modularity can be found in **Chapter 2.6**.

$$\boldsymbol{Mod_{edge}} = \frac{\boldsymbol{a_{edges}}}{\boldsymbol{A_{edges}}} \tag{2.4}$$

2.5 Environments

In this work, the simulations exposed each individual to sets containing 2, 4 or 8 environments. Besides that, the environments could present overlap (**OV**) or no overlap (**NO**) between environmental stimuli within the set. When the environments presented overlap os stimuli, we also considered that they would present overlap between the optimal output nodes.

As stated before, the environments are described as the list of input nodes (**Table 2.1 and 2.2**) that are activated when an individual is exposed to it. The overlapped sets have common stimuli for different environments within the set. The non-overlapped sets do not have common stimuli within the set. The output nodes that are expected to be activated to obtain a perfect fit to the set have code as the input nodes. In every environment, I_1 and I_2 are always activated and O_1 and O_2 are always expected to have been activated.

Tabela 2.1: Binary code of the input and output nodes of each environment of the overlapped sets.

		Nodes																	
Env Set	**Env**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
2OV	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
	2	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1
4OV	**1**	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	2	1	1	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0
	3	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
	4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1
8OV	**1**	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
	3	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
	4	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
	6	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
	7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
	8	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Tabela 2.2: Binary code of the input and output nodes of each environment of the non-overlapped sets.

		Nodes																	
Env Set	**Env**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
2NO	**1**	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
	2	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
4NO	**1**	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	3	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
8NO	**1**	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	3	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	4	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
	6	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
	7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
	8	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

2.6 Fitness and Modularity example

Taking the environment set with 4 non-overlapped environments, **4NO**, and the generic network **G1** (see **Figure 2.1**) as an example, it is possible to calculate the network's Fitness, Mod_{nodes} and Mod_{edges}.

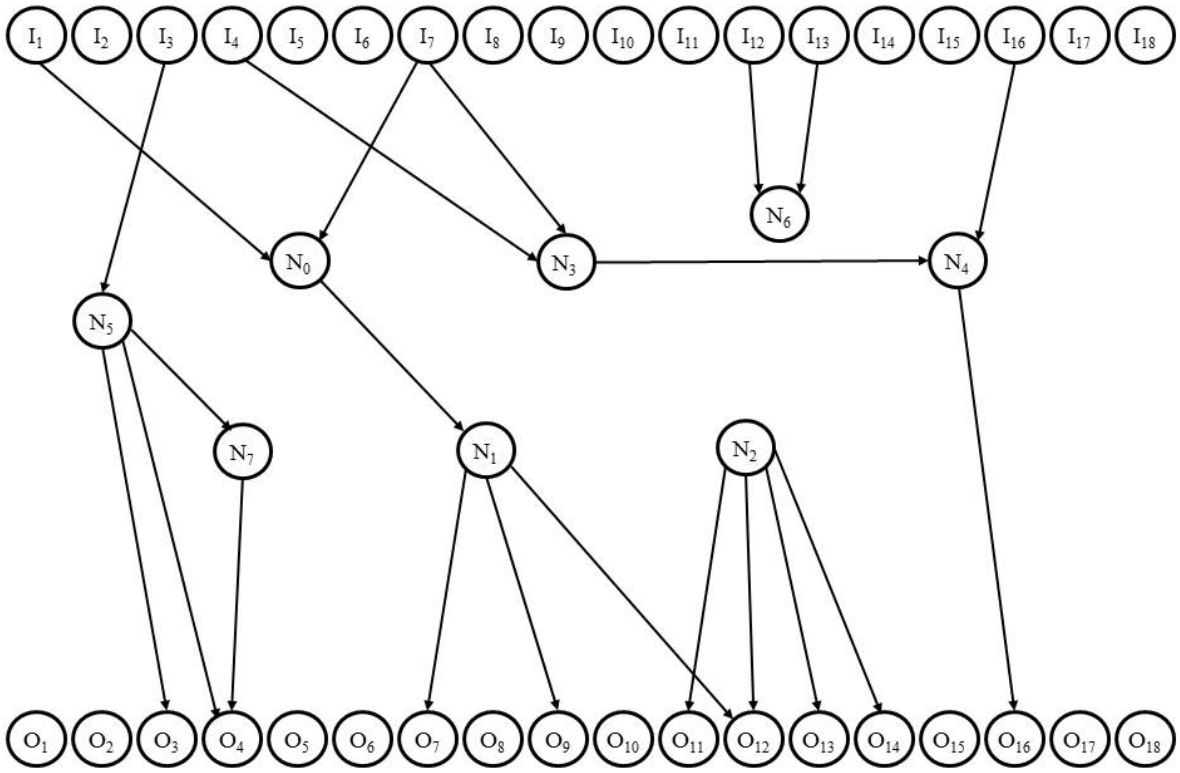

Figura 2.1: Generic network **G1** represented as a direct graph.I_n are the imnput nodes, N_n are the intermediate nodes and O_n are the output nodes. The arrows are directed edges and define the direction of propagation of the stimuli.

Environment 1 node activation paths:

{I_1, N_0, N_1, O_7, O_9, O_{12}}, {I_2}, {I_3, N_5, O_3, O_4}, {I_3, N_5, N_7, O_4}, {I_4, N_3, N_4, O_{16}}

Environment 2 node activation paths:

{I_1, N_0, N_1, O_7, O_9, O_{12}}, {I_2}, {I_7, N_0, N_1, O_7, O_9, O_{12}}, {I_7, N_3, N_4, O_{16}}, {I_8}, {I_9}, {I_{10}}

Environment 3 node activation paths:

$\{I_1, N_0, N_1, O_7, O_9, O_{12}\}$, $\{I_2\}$, $\{I_{11}\}$, $\{I_{12}, N_6\}$, $\{I_{13}, N_6\}$, $\{I_{14}\}$

Environment 4 node action paths:

$\{I_1, N_0, N_1, O_7, O_9, O_{12}\}$, $\{I_2\}$, $\{I_{15}\}$, $\{I_{16}, N_4, O_{16}\}$, $\{I_{17}\}$, $\{I_{18}\}$

So, the Output pattern of this network, for the different environments of this set is showed in **Table 2.3**.

Tabela 2.3: G1 output pattern. Each row represents a different enviroment to which the population was exposed to (environments 1, 2, 3 and 4 of the 4NO environment set). The 0 state means that the output node was not activated after the stimuli by the input nodes. The 1 state means that the signal traveled all the way from the input nodes to the output node, that became active.

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	O_{16}	O_{17}	O_{18}
0	0	1	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0
0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0
0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0

Tabela 2.4: G1 ideal output pattern. Each row represents a different enviroment to which the population was exposed to (environments 1, 2, 3 and 4 of the 4NO environment set). The 0 state means that the output node is not expected to be activated after the stimuli by the input nodes. The 1 state means that the signal should have traveled all the way from the input nodes to the output node and activate it. This pattern is the one that corresphonds to a maximum partial fitness in an environment of the set.

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	O_{16}	O_{17}	O_{18}
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

Therefore, the network's fitness is calculated as shown in **Equation 2.5**

$$\boldsymbol{Fit} = \left(\frac{\mathbf{2}}{\mathbf{10}}\right)^{8/4} \cdot \left(\frac{\mathbf{2}}{\mathbf{8}}\right)^{8/4} \cdot \left(\frac{\mathbf{1}}{\mathbf{8}}\right)^{8/4} \cdot \left(\frac{\mathbf{1}}{\mathbf{9}}\right)^{8/4} = \mathbf{4.82 \times 10^{-7}} \tag{2.5}$$

In terms of $\boldsymbol{Mod_{node}}$, looking at the generic network G1, we see that the nodes activated in any environments were:
$\{N_1, N_3, N_4, N_5, N_6, N_7\}$

and the nodes activated only in a single environment were:

{N_5, N_6}

Therefore, $\boldsymbol{Mod_{node}}$ value can be calculated as shown in **Equation 2.6**.

$$\boldsymbol{Mod_{node} = \frac{2}{6} = 0.333(3)} \tag{2.6}$$

As for $\boldsymbol{Mod_{edge}}$, looking at the generic network G1, we see that the edges activated in any environments were:

{ (I1,N0), (N0, N1), (N1, O7), (N1, O9), (N1, O12), (I3, N5), (N5, O3), (N5, O4), (N5, N7), (N7, O4), (I4, N3), (N3, N4), (N4, O16), (I7, N0), (I7, N3), (I12, N6), (I13, N6), (I16, N4) }

and the edges activated only in a single environment were:

{ (I3, N5), (N5, O3), (N5, O4), (N5, N7), (I4, N3), (I7, N0), (I7, N3), (I12, N6), (I13, N6), (I16, N4) }

Therefore, value can be calculated as shown in **Equation 2.7**.

$$\boldsymbol{Mod_{edge} = \frac{10}{18} = 0.555(5)} \tag{2.7}$$

2.7 Mutational Parameters

Besides varying both the number and overlap condition of the environments, we also varied the mutational parameters: edge mutation probability ($\boldsymbol{prob_{mut}}$), node duplication probability ($\boldsymbol{prob_{duplic}}$) and node elimination probability ($\boldsymbol{prob_{elim}}$). The variations were made between 4 possible values: 0.01, 0.005, 0.001 and 0.0001. A list of all the essays performed can be found in **Table 2.5**). A spreadsheet containing the final output values of the last generations can be found in the digital version of this dissertation's annex (Annex/Resume.xlsx).

Tabela 2.5: Mutational parameters combinations ($prob_{duplic}$, $prob_{mut}$ and $prob_{elim}$). The variations were made between 4 possible values: 0.01, 0.005, 0.001 and 0.0001

ID	$\boldsymbol{prob_{duplic}}$	$\boldsymbol{prob_{mut}}$	$\boldsymbol{prob_{elim}}$
001all	0.001	0.001	0.001
005all	0.005	0.005	0.005
01all	0.01	0.01	0.01
001duplic0001	0.0001	0.001	0.001
001mut0001	0.001	0.0001	0.001
001elim0001	0.001	0.001	0.0001
001duplic005	0.005	0.001	0.001
001mut005	0.001	0.005	0.001
001elim005	0.001	0.001	0.005
005duplic001	0.001	0.005	0.005
005mut001	0.005	0.001	0.005
005elim001	0.005	0.005	0.001
005duplic01	0.01	0.005	0.005
005mut01	0.005	0.01	0.005
005elim01	0.005	0.005	0.01

As for the networks' evolution dynamic, as stated before, each individual's network starts in the first generation with 18 fixed input and output nodes and only 1 unconnected intermediate node. The intermediate nodes number can fluctuate in each generation by duplication or elimination mechanisms. These can be applied to any intermediate node, in a given generation, with a probability of $\boldsymbol{prob_{duplic}}$ and it's made a copy of the target node, maintaining every connection of the original node (see **Figure 2.2**). The elimination process occurs in any node with a probability of $\boldsymbol{prob_{elim}}$ and the target node is removed from the graph and all of its connections as well (see **Figure 2.4**). The edge mutation process that can occur on any existing edge of the graph being removed with a probability of $\boldsymbol{prob_{mut}}$, and any possible connection that two nodes can theoretically form, can be added to the graph with, also, a probability of $\boldsymbol{prob_{mut}}$ (see **Figure 2.3**).

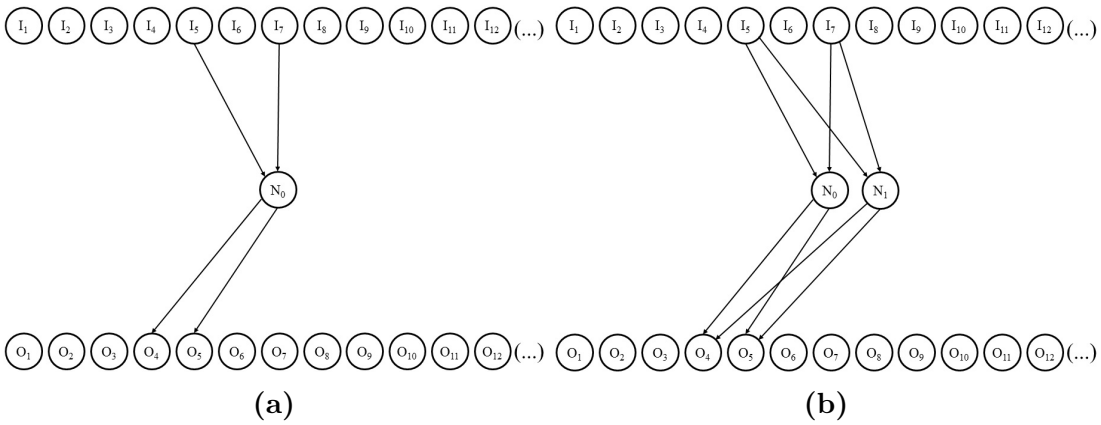

Figura 2.2: Node duplication process on generic network **G2**. Sequencial process starting in the network's initial state (a), where $\boldsymbol{N_0}$ has two input edges, from $\boldsymbol{I_5}$ and $\boldsymbol{I_7}$, and two output edges, to $\boldsymbol{O_4}$ and $\boldsymbol{O_5}$. From (a) to (b) $\boldsymbol{N_0}$ is duplicated, creating a new node, $\boldsymbol{N_1}$, that inherits the same input and output edges as the original node, creating the generic network **G3**.

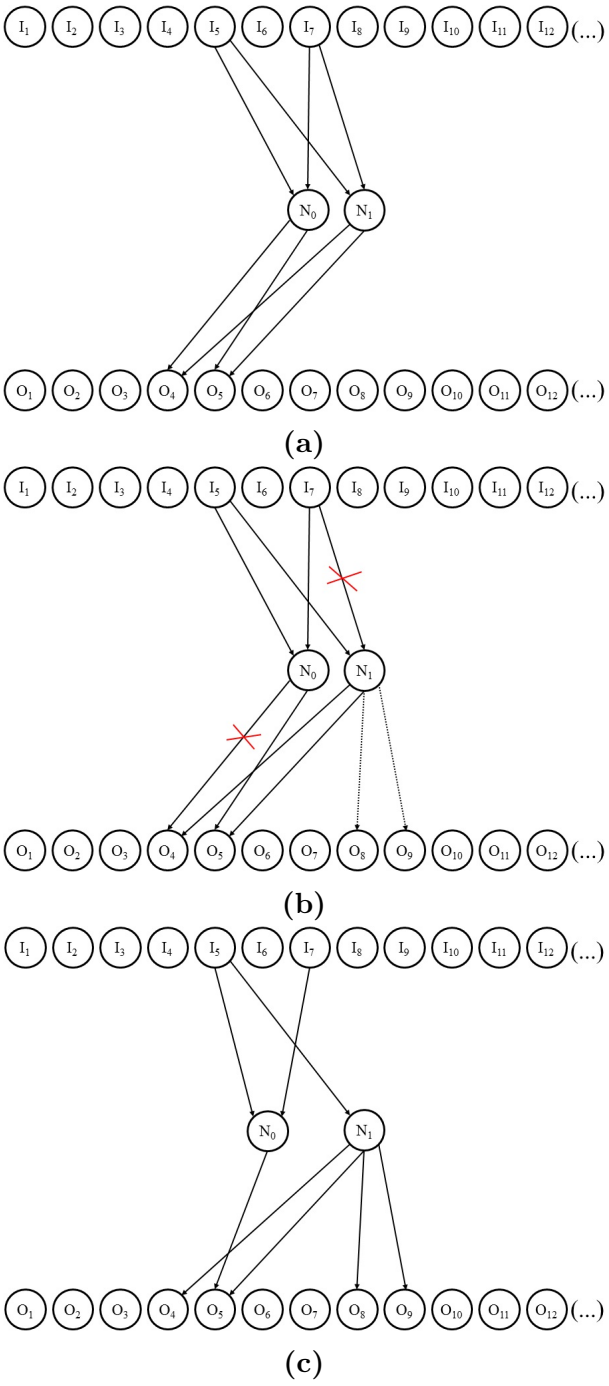

Figura 2.3: Edge mutation process on generic network **G3**. Sequencial process starting in the network's initial state (a), where $\boldsymbol{N_0}$ and $\boldsymbol{N_1}$ have two input edges, from $\boldsymbol{I_5}$ and $\boldsymbol{I_7}$, and two output edges, to $\boldsymbol{O_4}$ and $\boldsymbol{O_5}$. From (a) to (c), the edge mutation process, represented in (b), eliminated the edge from $\boldsymbol{I_7}$ to $\boldsymbol{N_1}$ and the one from $\boldsymbol{N_0}$ to $\boldsymbol{O_4}$. It also added an edge from $\boldsymbol{N_1}$ to $\boldsymbol{O_8}$ and other on from $\boldsymbol{N_1}$ to $\boldsymbol{O_9}$, creating the generic network **G4**.

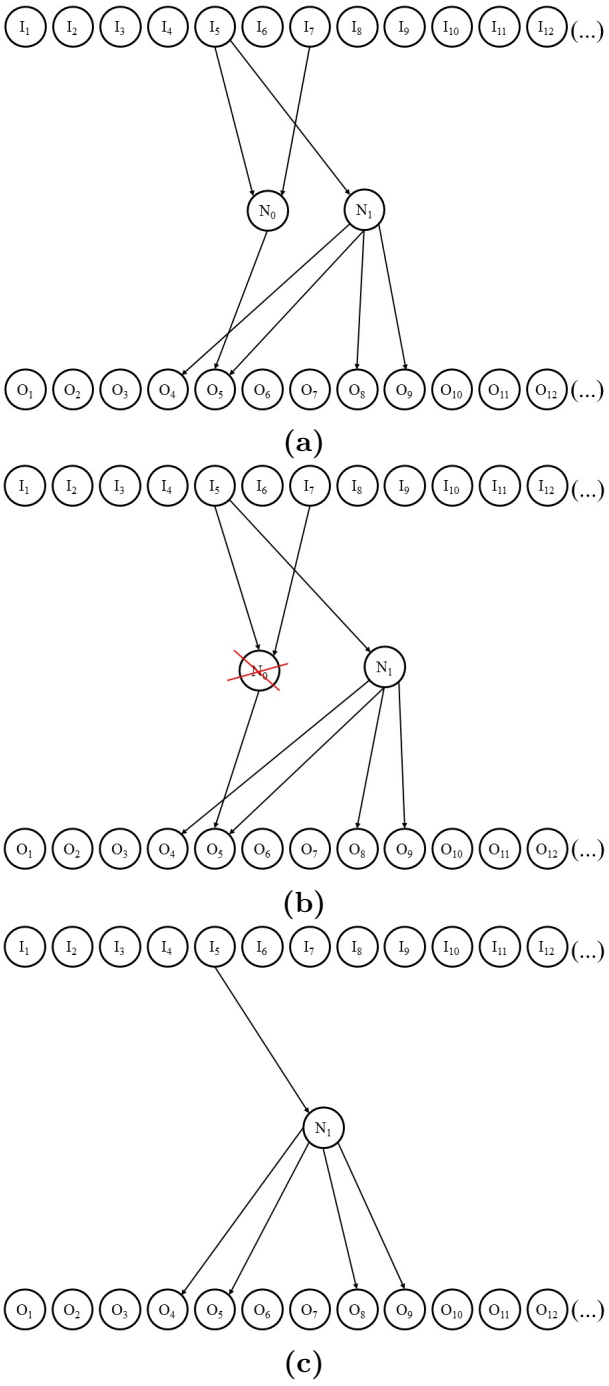

Figura 2.4: Elimination process on generic network **G4**. Sequencial process starting in the network's initial state (a), where $\boldsymbol{N_0}$ has two input edges, from $\boldsymbol{I_5}$ and $\boldsymbol{I_7}$, and one output edges, to $\boldsymbol{O_4}$ and $\boldsymbol{O_5}$, and $\boldsymbol{N_1}$ has one input edge from $\boldsymbol{I_5}$ and four output edges to $\boldsymbol{O_4}$, $\boldsymbol{O_5}$, $\boldsymbol{O_8}$ and $\boldsymbol{O_9}$. From (a) to (c), the node elimination process, represented in (b), eliminated the node $\boldsymbol{N_0}$ and, as a consequence, every input and output node associated to it.

2.8 Statistical Analysis

Each simulation with a different combination of number of environments, overlap condition and parameters values was replicated 5 times.

For each generation, in each simulation, the mean (of the 1000 individuals) values of fitness, number of nodes, number of edges, modularity based on nodes and modularity based on edges was calculated.

In the parameter variation analysis on fitness and modularity it was applied a linear regression, being tested if the slope was significantly different than zero.

To compare the fitness and modularity mean values of the simulations' last generation with the number of environments and their overlap state of different environments, it was used a 2-way ANOVA based in permutations [17] since the different replica groups presented heterogeneous dispersion.

The analysis of the output data of the evolution model was made using Python expanded with the Numpy and MatPlotLib modules and the statistical programming language R.

2.9 Model Implementation

The simulation model was built in Python 2.7 programming language using some expansion modules (Pickle, Numpy and Scipy). The updated version of the code used in this work can be found in the digital version of this dissertation's annex, or in my personal github repository: https://github.com/danvilar/mygenes.

Capítulo 3

Results

3.1 Impact of Parameter Variation

To evaluate the influence of the different types of mutation on the evolution of the networks' fitness and modularity, the simulations were made varying each mutational parameter individually around two reference states:

- $\boldsymbol{prob_{duplic} = prob_{mut} = prob_{elim} = 0.005}$
- $\boldsymbol{prob_{duplic} = prob_{mut} = prob_{elim} = 0.001}$

3.1.1 $\boldsymbol{prob_{duplic}}$ Variation

For any of the reference states, a higher value of $\boldsymbol{prob_{duplic}}$ leads to a significant higher fitness and modularity values when the individuals are exposed to 8 environments without overlap.

For the 0.001 reference state (**Table 3.1**), a higher value of $\boldsymbol{prob_{duplic}}$ lead to a significantly higher fitness and modularity values when the individuals are exposed to 8 environments with overlap. Besides that, in this reference state, a higher value of $\boldsymbol{prob_{duplic}}$ leads to a significant lower value of fitness when the individuals are exposed to 2 environments without overlap, but it does not significantly affect the modularity values.

These variations suggest that a higher $\boldsymbol{prob_{duplic}}$ of the nodes makes it easier for the higher fitness and modularity networks to be selected in more complex environments. In simpler environments, like the 2 environments with overlap, and lower mutation rates (0.001 reference state) an over duplication can prevent the selection of the optimal networks.

Tabela 3.1: $\boldsymbol{prob_{duplic}}$ impact on the output parameters ($\boldsymbol{Fitness}$, $\boldsymbol{Mod_{node}}$ and $\boldsymbol{Mod_{edge}}$). $\boldsymbol{prob_{duplic}}$ values: 0.0001, 0.001 and 0.005) $\boldsymbol{prob_{mut}}$ and $\boldsymbol{prob_{elim}}$ fixed to 0.001. Fitness significant on the environment sets (Env) 2OV, 8OV and 8NO. Modularities significant on Env 8OV and 8NO

Env	$\boldsymbol{Fitness}$		$\boldsymbol{Mod_{node}}$		$\boldsymbol{Mod_{edge}}$	
	slope	p value	slope	p-value	slope	p-value
2OV	1.83E+01	6.64E-01	1.98E+01	6.05E-01	2.37E+01	4.28E-01
2NO	-2.21E+00	3.40E-02	-4.37E-01	1.25E-01	-3.55E-01	1.19E-01
4OV	5.37E+01	1.12E-01	4.17E+01	1.95E-01	4.11E+01	1.18E-01
4NO	6.42E+01	1.03E-01	5.04E+01	1.39E-01	2.40E+01	9.63E-02
8OV	9.78E+01	2.87E-03	1.02E+02	1.67E-02	6.23E+01	1.37E-02
8NO	6.66E+01	1.26E-03	1.47E+02	2.32E-04	3.56E+01	1.30E-04

For the 0.005 reference state (**Table 3.2**), we can see that the only environment set that has a significant impact is the 8 NO, where the fitness and modularity both have a tendency to be higher with higher duplication rate. Empirically, organisms in the 8 environment sets will need more nodes to start showing fitness. The non-overlapped variant may not need that much node because the unique phenotype node that are needed for an individual to survive are less when comparing to the overlapped variant, therefore needing more nodes to increase the chance of activating those nodes.

Tabela 3.2: $\boldsymbol{prob_{duplic}}$ impact on the output parameters ($\boldsymbol{Fitness}$, $\boldsymbol{Mod_{node}}$ and $\boldsymbol{Mod_{edge}}$). $\boldsymbol{prob_{duplic}}$ values: 0.001, 0.005 and 0.01) $\boldsymbol{prob_{mut}}$ and $\boldsymbol{prob_{elim}}$ fixed to 0.005. Significant only on environment set 8NO.

Env	$\boldsymbol{Fitness}$		$\boldsymbol{Mod_{node}}$		$\boldsymbol{Mod_{edge}}$	
	slope	p value	slope	p-value	slope	p-value
2OV	-2.03E+01	1.81E-01	-2.09E+01	1.43E-01	-1.43E+01	2.01E-01
2NO	-1.26E+00	2.04E-01	-9.70E-03	9.80E-01	-2.49E-02	9.25E-01
4OV	-9.65E+00	1.22E-01	-1.62E+01	2.10E-01	-1.19E+01	2.12E-01
4NO	-1.11E+00	4.06E-01	1.74E-01	7.21E-01	2.17E-01	5.52E-01
8OV	6.31E-01	1.47E-01	1.33E+01	1.00E-01	1.21E+01	7.64E-02
8NO	2.20E+00	1.41E-03	6.35E+01	1.14E-04	1.05E+01	1.16E-04

3.1.2 $\boldsymbol{prob_{mut}}$ Variation

Globally, we see that higher value of $\boldsymbol{prob_{mut}}$ leads to lower fitness and modularity values. This effect is more visible in the reference state 0.005 (**Table 3.4**). In the reference state 0.001 (**Table 3.3**), the $\boldsymbol{prob_{mut}}$ value is already too low, and this parameter's variation only affects the organisms exposed to 8 environments with overlap. Really high $\boldsymbol{prob_{mut}}$ values will make the descendants of individuals who originally have high fitness values lose, by mutation, important connections between nodes that were the ones making their

parent individual successful.

Tabela 3.3: $\boldsymbol{prob_{mut}}$ impact on the output parameters ($\boldsymbol{Fitness}$, $\boldsymbol{Mod_{node}}$ and $\boldsymbol{Mod_{edge}}$). $\boldsymbol{prob_{mut}}$ values: 0.0001, 0.001 and 0.005) $\boldsymbol{prob_{duplic}}$ and $\boldsymbol{prob_{elim}}$ fixed to 0.005. Significant only on environment set 8OV

Env	$\boldsymbol{Fitness}$		$\boldsymbol{Mod_{node}}$		$\boldsymbol{Mod_{edge}}$	
	slope	p value	slope	p-value	slope	p-value
2OV	-6.70E+01	1.53E-01	-7.41E+01	1.18E-01	-5.47E+01	1.56E-01
2NO	6.56E+00	8.72E-01	3.01E+01	3.78E-01	2.04E+01	3.80E-01
4OV	-4.59E+01	2.24E-01	3.08E+01	2.33E-01	1.37E+01	5.28E-01
4NO	-3.12E+01	4.08E-01	-8.15E-01	9.54E-01	-5.96E+00	3.90E-01
8OV	-6.96E+01	1.47E-02	-7.50E+01	2.01E-02	-7.14E+01	9.65E-03
8NO	-2.33E+00	2.20E-01	-5.44E+01	1.49E-01	-1.61E+01	4.28E-02

Tabela 3.4: $\boldsymbol{prob_{mut}}$ impact on the output parameters ($\boldsymbol{Fitness}$, $\boldsymbol{Mod_{node}}$ and $\boldsymbol{Mod_{edge}}$). $\boldsymbol{prob_{mut}}$ values: 0.001, 0.005 and 0.01) $\boldsymbol{prob_{duplic}}$ and $\boldsymbol{prob_{elim}}$ fixed to 0.005. Significant on all environment set excpet 2OV.

Env	$\boldsymbol{Fitness}$		$\boldsymbol{Mod_{node}}$		$\boldsymbol{Mod_{edge}}$	
	slope	p value	slope	p-value	slope	p-value
2OV	-1.55E+01	3.77E-01	2.69E+00	8.80E-01	2.75E+00	8.51E-01
2NO	-4.28E+01	5.29E-15	-1.19E+01	2.10E-13	-8.29E+00	1.36E-13
4OV	-8.17E+01	9.79E-11	-7.69E+01	5.06E-04	-6.46E+01	4.11E-04
4NO	-6.72E+01	1.45E-11	-1.91E+01	4.95E-04	-1.51E+01	4.22E-08
8OV	-6.80E+01	4.87E-05	-9.35E+01	2.37E-07	-7.54E+01	1.28E-06
8NO	-1.55E+01	2.30E-02	-5.97E+01	8.13E-03	-1.72E+01	5.34E-04

3.1.3 $\boldsymbol{prob_{elim}}$ Variation

As we can see from the data presented in **Table 3.5** and **Table 3.6**, the variation of this parameter does not seem to have any significant effect on either fitness or modularity values of the tested organisms under any environmental conditions. The explanation for this observation may be that the elimination effect can be replaced with the negative selection of networks with unwanted nodes. The edges mutation and node duplication facilitate the arise of new structures that can contribute in a positive or negative way to the fitness values. On the other hand, the elimination just deletes structures that already exist in the population.

Tabela 3.5: $\boldsymbol{prob_{elim}}$ impact on the output parameters ($\boldsymbol{Fitness}$, $\boldsymbol{Mod_{node}}$ and $\boldsymbol{Mod_{edge}}$). $\boldsymbol{prob_{elim}}$ values: 0.0001, 0.001 and 0.005) $\boldsymbol{prob_{duplic}}$ and $\boldsymbol{prob_{mut}}$ fixed to 0.001. Not significant on any environment set.

Env	$\boldsymbol{Fitness}$		$\boldsymbol{Mod_{node}}$		$\boldsymbol{Mod_{edge}}$	
	slope	p value	slope	p-value	slope	p-value
2OV	-4.72E+01	2.78E-01	-2.05E+01	6.38E-01	-1.14E+01	7.42E-01
2NO	-7.99E-01	3.81E-01	1.40E-02	9.60E-01	5.54E-04	9.98E-01
4OV	-3.02E+00	5.51E-02	5.89E+00	5.08E-01	4.64E+00	5.05E-01
4NO	-1.40E+00	3.54E-01	1.75E+01	8.74E-02	5.92E+00	8.71E-02
8OV	-3.06E+00	8.91E-01	1.45E+01	2.66E-01	6.60E+00	5.80E-01
8NO	-8.19E-01	6.57E-01	1.45E+01	7.48E-01	2.87E+00	7.85E-01

Tabela 3.6: $\boldsymbol{prob_{elim}}$ impact on the output parameters ($\boldsymbol{Fitness}$, $\boldsymbol{Mod_{node}}$ and $\boldsymbol{Mod_{edge}}$). $\boldsymbol{prob_{elim}}$ values: 0.001, 0.005 and 0.01) $\boldsymbol{prob_{duplic}}$ and $\boldsymbol{prob_{mut}}$ fixed to 0.005. Not significant on any environment set.

Env	$\boldsymbol{Fitness}$		$\boldsymbol{Mod_{node}}$		$\boldsymbol{Mod_{edge}}$	
	slope	p value	slope	p-value	slope	p-value
2OV	-1.71E+00	9.12E-01	3.63E+00	8.04E-01	-9.28E-02	9.94E-01
2NO	2.38E-01	8.06E-01	3.94E-01	4.16E-01	4.16E-01	2.04E-01
4OV	-2.21E+00	9.05E-02	-6.72E-01	2.33E-01	-7.91E-01	1.07E-01
4NO	-1.16E+00	2.58E-01	1.13E-01	7.56E-01	6.94E-02	7.93E-01
8OV	-3.21E-01	6.53E-01	-6.73E+00	3.05E-01	-5.75E+00	2.29E-01
8NO	-4.00E-01	2.94E-01	-1.56E+01	4.17E-01	-2.11E+00	5.32E-01

In this set of variations of the three parameters, when there is a significant impact, there is a concordance in the fitness and modularity (both node and edge based) variation sign. This suggests that the architectures with better fitness tend to be, simultaneously, modular networks.

3.2 Number of Environments and Overlap Impact

3.2.1 Impact on Final Population Characteristics

The previous study allowed us to identify the combination of parameters $Prob_{duplic} = 0.005$, $Prob_{mut} = 0.001$ and $Prob_{elim} = 0.001$ as the one that leads to better fitness values within the tested environment sets. Based on this combination, we studied the number of environments impact and the overlap state presence between the stimuli of each environment on the fitness and signalling network architecture.

In terms of fitness, the 8 environments make the networks' evolution harder to reach high fitness results. The existence of overlap, in this case, allows the network to achieve higher

fitness results but not a significantly higher modularity (**Figures 3.1-3.3**). On the other hand, with just 2 environments, the overlap prevents the network to evolve in order to achieve higher fitness values. This opposite effect between the 2 and 8 environments leads to a significant interaction between the two factors, number of environments and overlap state, in the 2way-ANOVA analysis ($p = 0.001$). The overlap does not seem to have any impact with 4 environments.

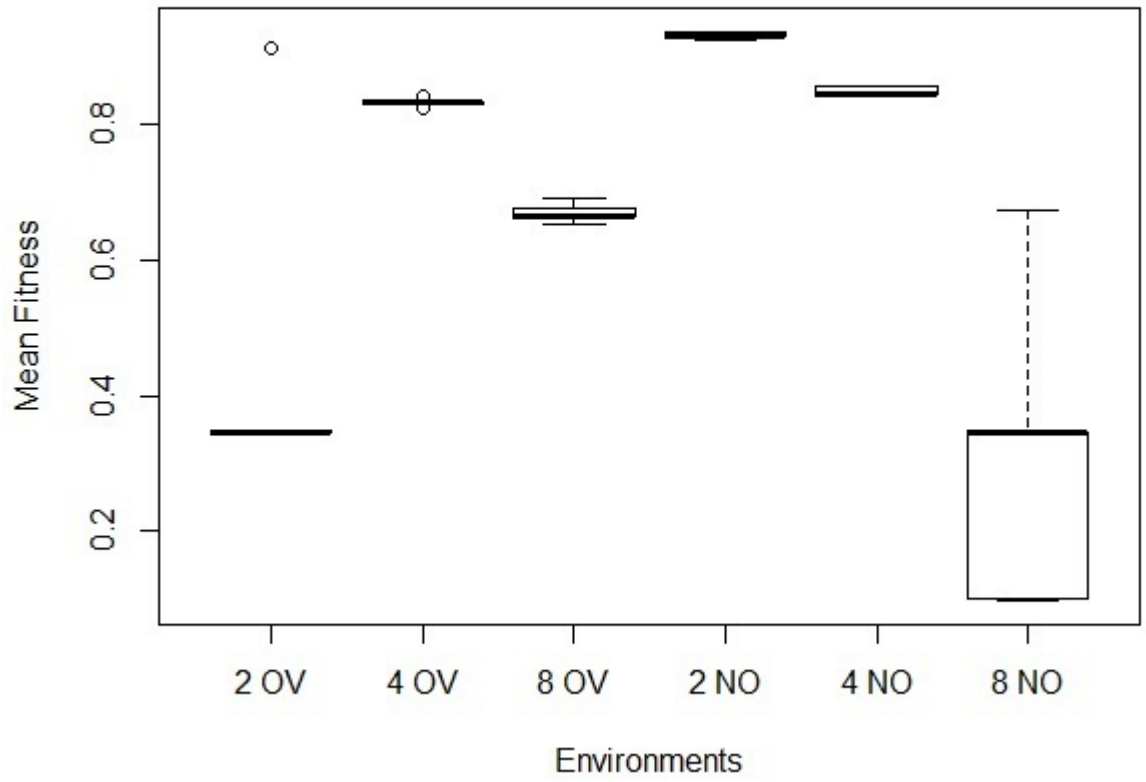

Figura 3.1: Boxplot of the effect and interaction between the number of environments and overlap state on ***Fitness***. In the 2way-ANOVA analysis the number of environments is significant ($p = 0.001$) and the interaction between the number of environments and the overalp state is significat as well ($p = 0.001$).

Both nodes and edge based modularity present a similar behaviour (**Figures 3.2 and 3.3**). The number of environments effect is not so visible, especially without overlap. With overlap, the most visible effect is the low modularity of the individuals exposed to 2 environments, which lead us again to the significant effect of the interaction between the number of environments and the overlap ($p = 0.002$ for Mod_{node} and $p = 0.003$ for Mod_{Edge}). The overlap's effect is significant ($p = 0.003$ for Mod_{node} and $p = 0.001$ for Mod_{Edge}), but that is mainly due to the low values on the essays with 2 overlapped environments. With 4 and 8 environments it does not seem to have any significant impact.

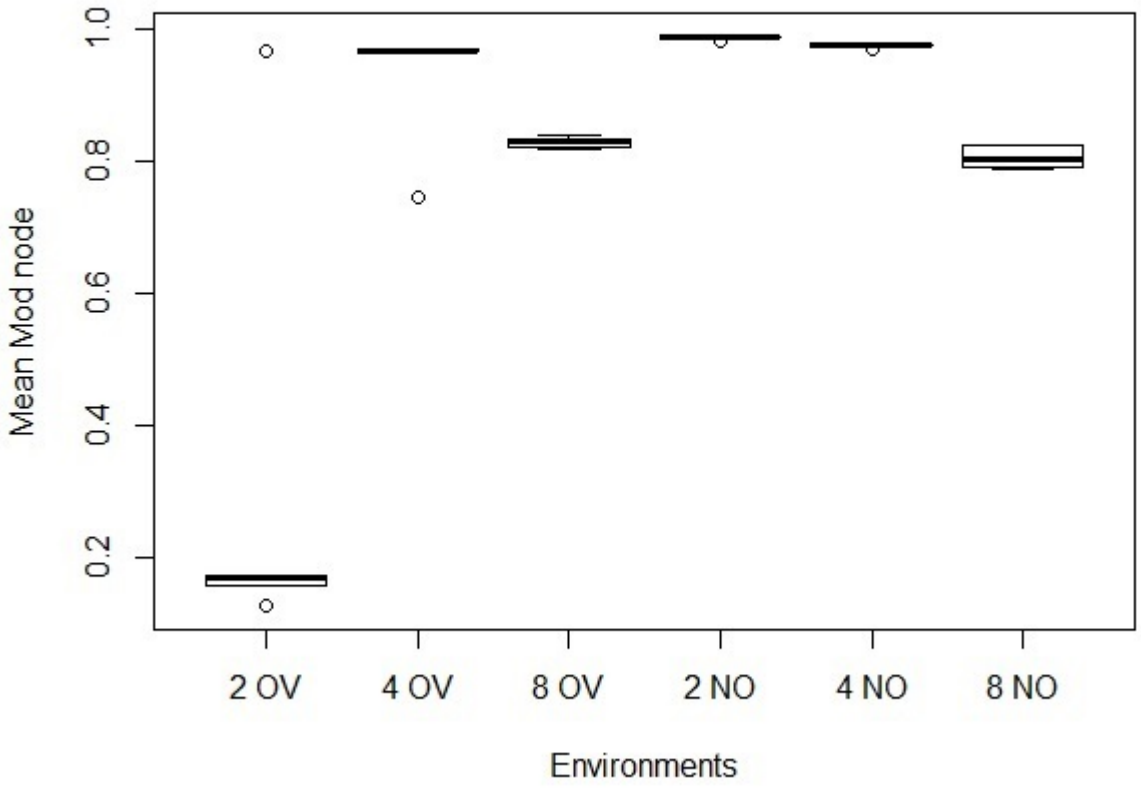

Figura 3.2: Boxplot of the effect and interaction between the number of environments and overlap state on $\boldsymbol{Mod_{node}}$. In the 2way-ANOVA analysis the overlap state is significant ($p = 0.004$) and the interaction between the number of environments and the overalp state is significat as well ($p = 0.003$).

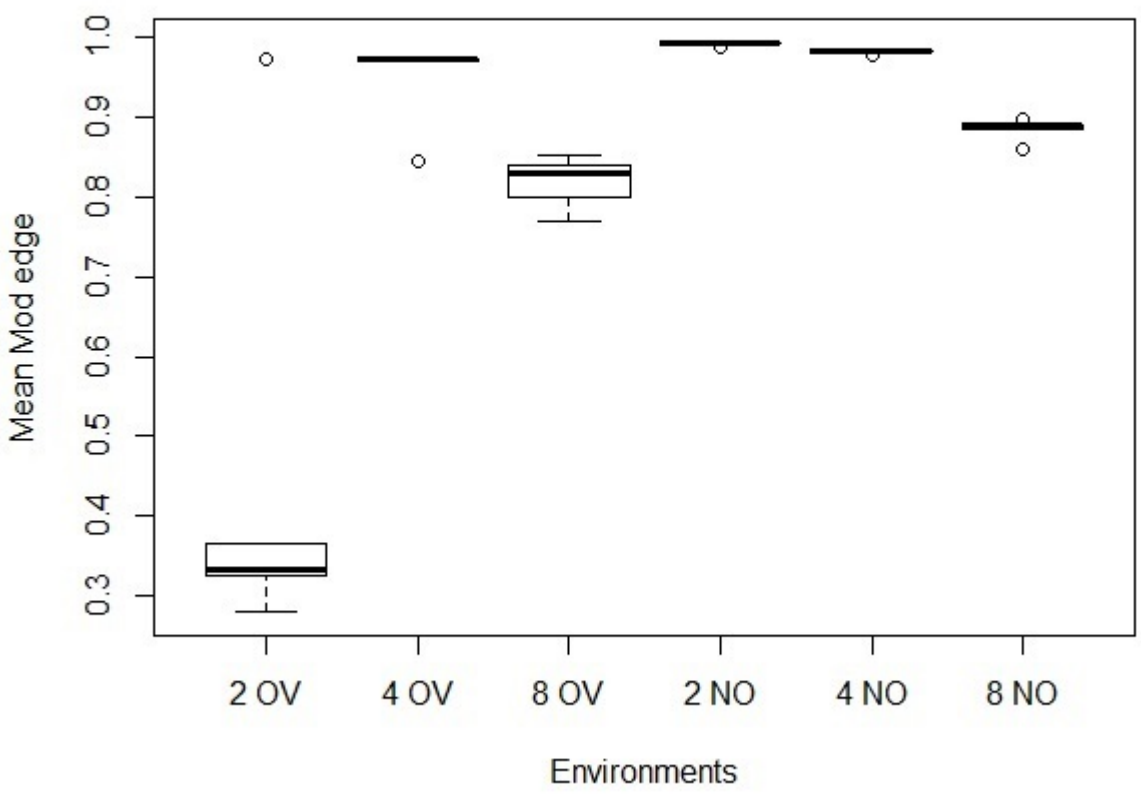

Figura 3.3: Boxplot of the effect and interaction between the number of environments and overlap state on $\boldsymbol{Mod_{edge}}$. In the 2way-ANOVA analysis the overlap state is significant ($p = 0.001$) and the interaction between the number of environments and the overalp state is significat as well ($p = 0.002$).

In terms of the number of nodes and edges (**Figures 3.4 and 3.5**), it is expectable a higher number along with the number of environments. This is visible for the number of nodes without overlap and partially for the edges, without overlap. With overlap, the main result is the much higher number of nodes and edges when the organisms are exposed to just 2 overlapped environments. This may justify the low fitness values under these conditions. With 4 and 8 overlapped environments, the higher number of environments lead to a higher number of nodes, as expected.

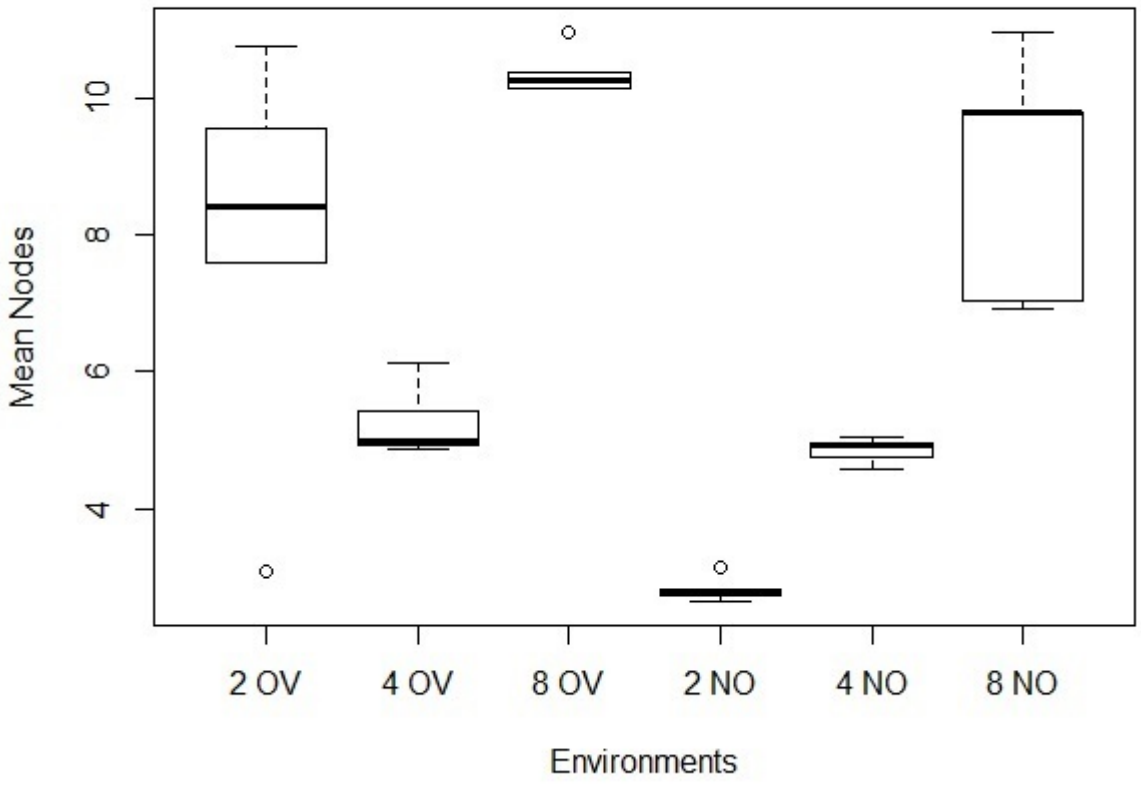

Figura 3.4: Boxplot of the effect and interaction between the number of environments and overlap state on ***Nodes***. In the 2way-ANOVA analysis the number of environments is significant ($p = 0.001$) and the overalp state is significat as well ($p = 0.002$).

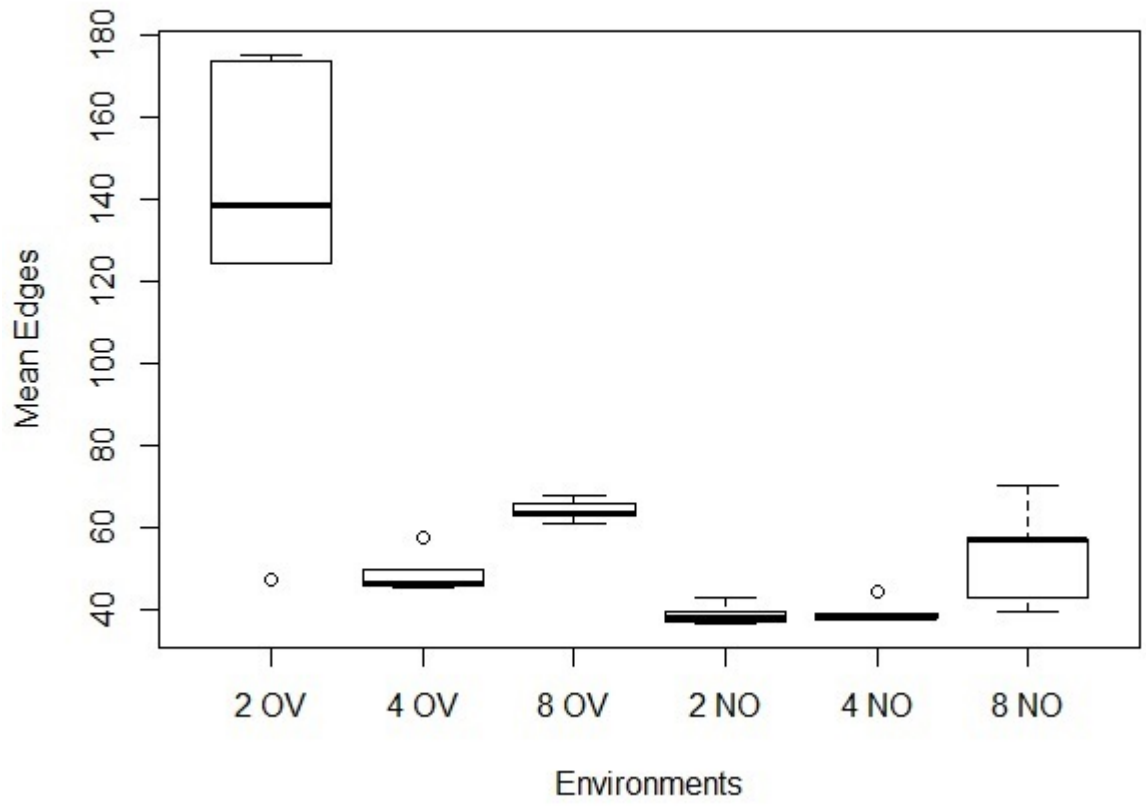

Figura 3.5: Boxplot of the effect and interaction between the number of environments and overlap state on the mean number of edges. In the 2way-ANOVA analysis the overlap state is significant ($p = 0.002$) and the interaction between the number of environments and the overalp state is significat as well ($p = 0.005$).

In this study, there are some interesting observations, such as the fact that with 8 environments, comparing the overlapped with the non-overlapped state, similar modularity values correspond to different fitness values (higher when overlapped). This means that the overlap makes the better adapted networks evolution easier, but this does not affect the networks' modularity.

3.2.2 The Effect of Environment and Overlap on Networks' Temporal Evolution Profile

Besides the effort to understand the environment number and overlap state impact in the final result of the evolutionary process, we also evaluated its dynamics monitoring the fitness mean and modularity mean variation for each successive generation (**Figures3.6-3.11**).

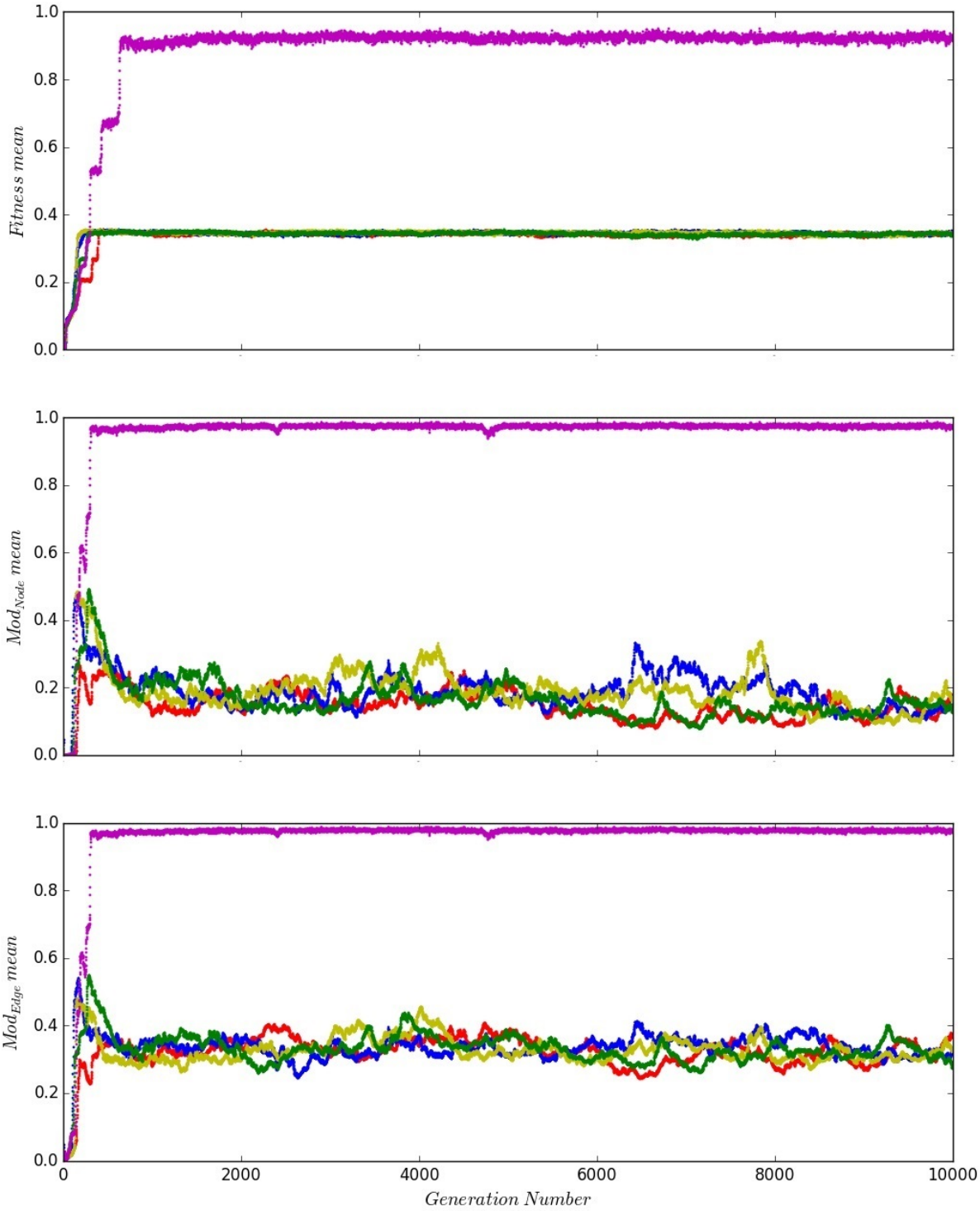

Figura 3.6: Fitness mean values, $\text{Mod}_{\boldsymbol{node}}$ mean values and $\text{Mod}_{\boldsymbol{edge}}$ mean values temporal evolution profile during 10000 generations. These mean values are calculated with the correspondent value of each one of the 1000 individuals in each generation. The input parameters of this essay were $\boldsymbol{prob_{duplic} = 0.005}$, $\boldsymbol{prob_{mut} = prob_{elim} = 0.001}$ and using the Environment Set **2OV**. Red dots: first replica. Blue dots: second replica. Yellow dots: third replica. Green dots: fourth replica. Purple dots: fifth replica.

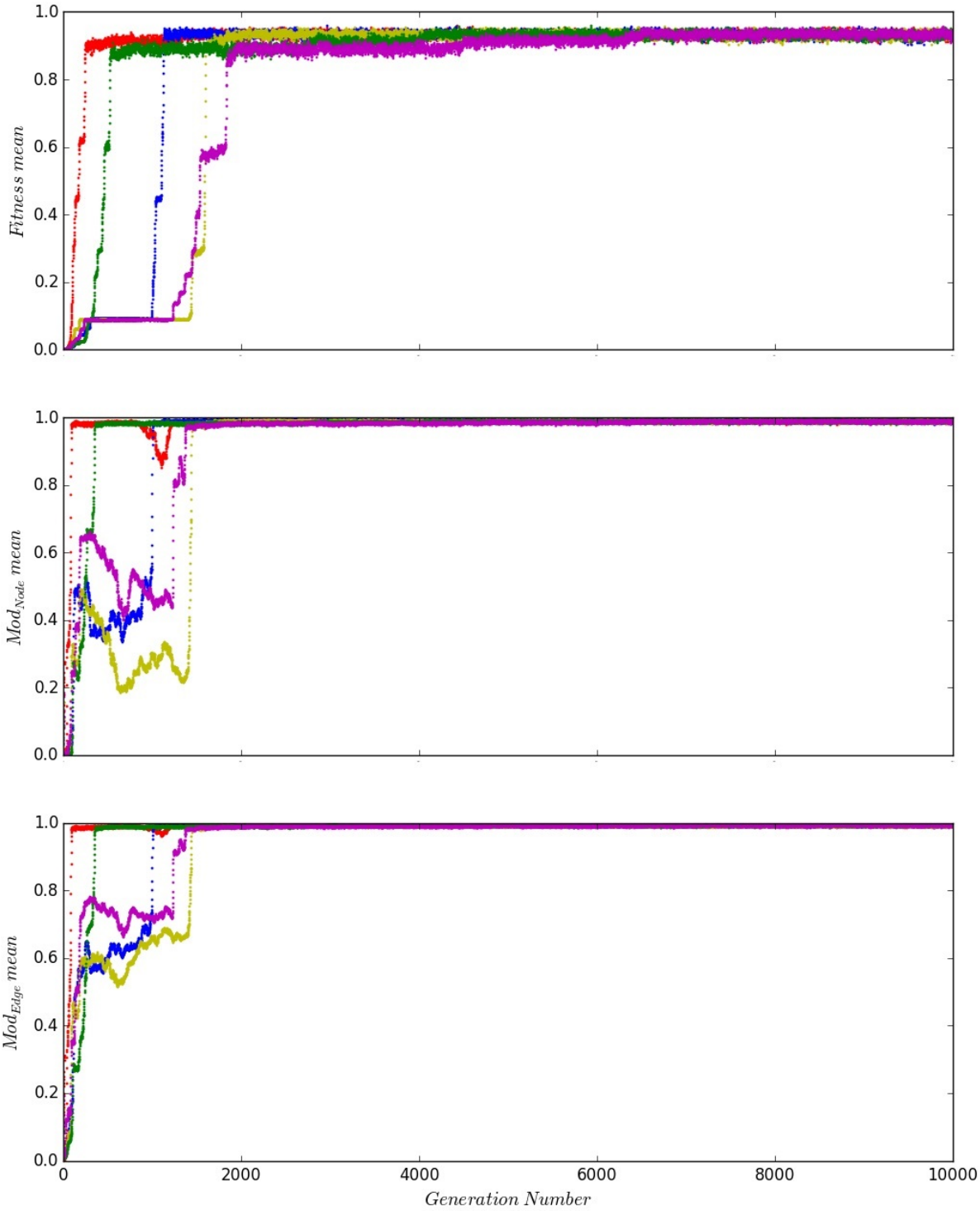

Figura 3.7: Fitness mean values, $\text{Mod}_{\boldsymbol{node}}$ mean values and $\text{Mod}_{\boldsymbol{edge}}$ mean values temporal evolution profile during 10000 generations. These mean values are calculated with the correspondent value of each one of the 1000 individuals in each generation. The input parameters of this essay were $\boldsymbol{prob_{duplic} = 0.005}$, $\boldsymbol{prob_{mut} = prob_{elim} = 0.001}$ and using the Environment Set **2NO**. Red dots: first replica. Blue dots: second replica. Yellow dots: third replica. Green dots: fourth replica. Purple dots: fifth replica.

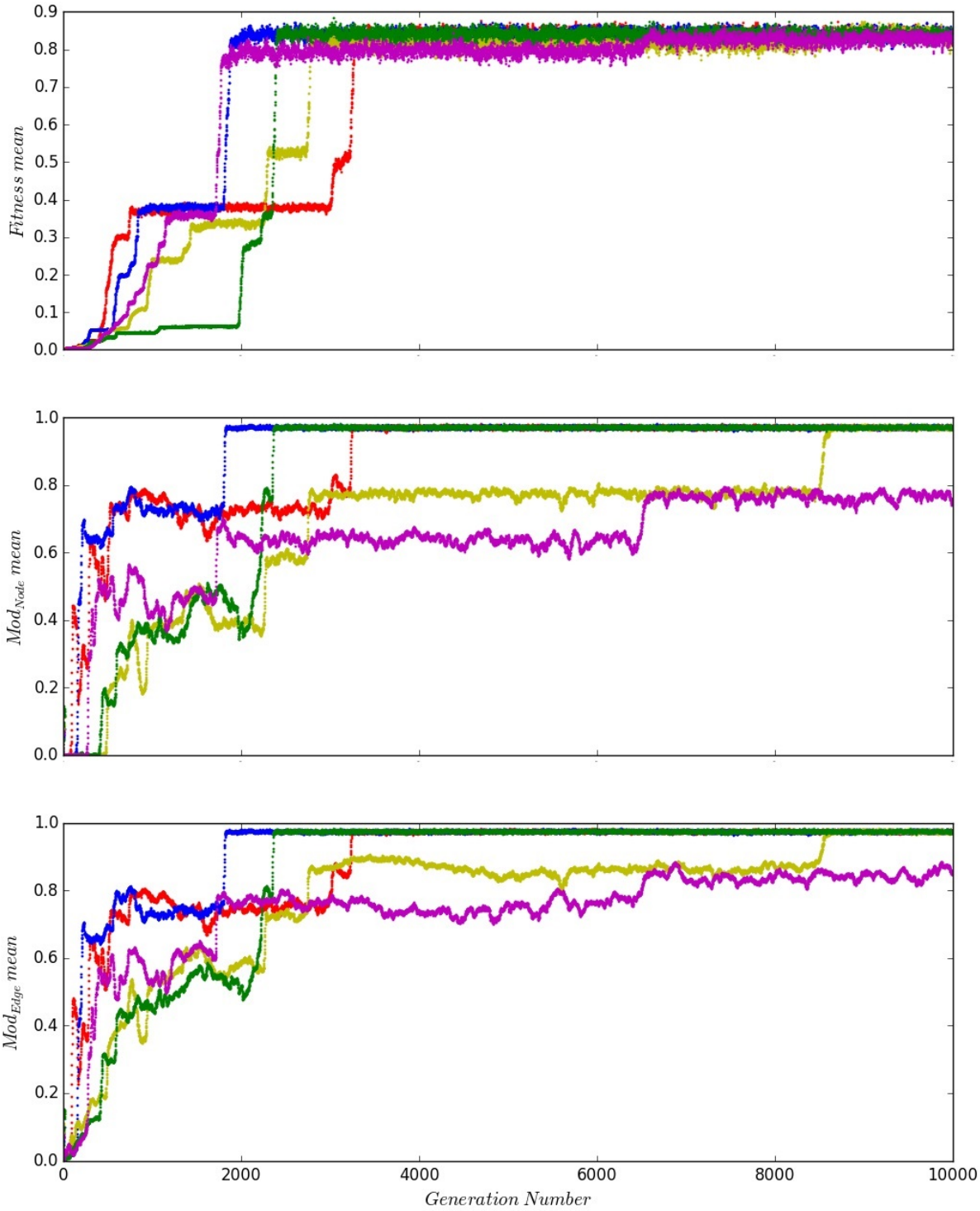

Figura 3.8: Fitness mean values, $\text{Mod}_{\boldsymbol{node}}$ mean values and $\text{Mod}_{\boldsymbol{edge}}$ mean values temporal evolution profile during 10000 generations. These mean values are calculated with the correspondent value of each one of the 1000 individuals in each generation. The input parameters of this essay were $\boldsymbol{prob_{duplic} = 0.005}$, $\boldsymbol{prob_{mut} = prob_{elim} = 0.001}$ and using the Environment Set **4OV**. Red dots: first replica. Blue dots: second replica. Yellow dots: third replica. Green dots: fourth replica. Purple dots: fifth replica.

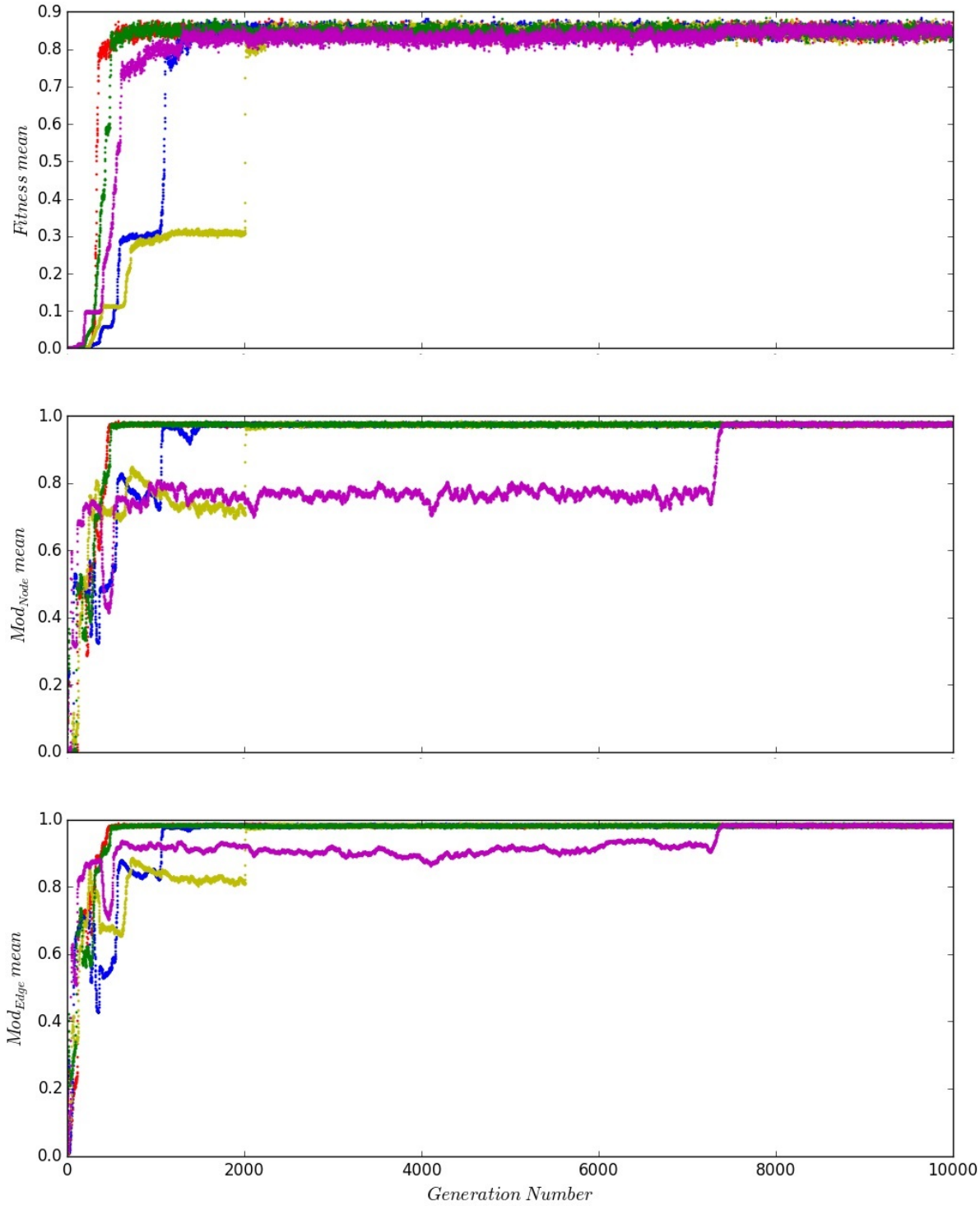

Figura 3.9: Fitness mean values, $\text{Mod}_{\boldsymbol{node}}$ mean values and $\text{Mod}_{\boldsymbol{edge}}$ mean values temporal evolution profile during 10000 generations. These mean values were calculated with the correspondent value of each one of the 1000 individuals in each generation. The input parameters of this essay are $\boldsymbol{prob_{duplic} = 0.005}$, $\boldsymbol{prob_{mut} = prob_{elim} = 0.001}$ and using the Environment Set **4NO**. Red dots: first replica. Blue dots: second replica. Yellow dots: third replica. Green dots: fourth replica. Purple dots: fifth replica.

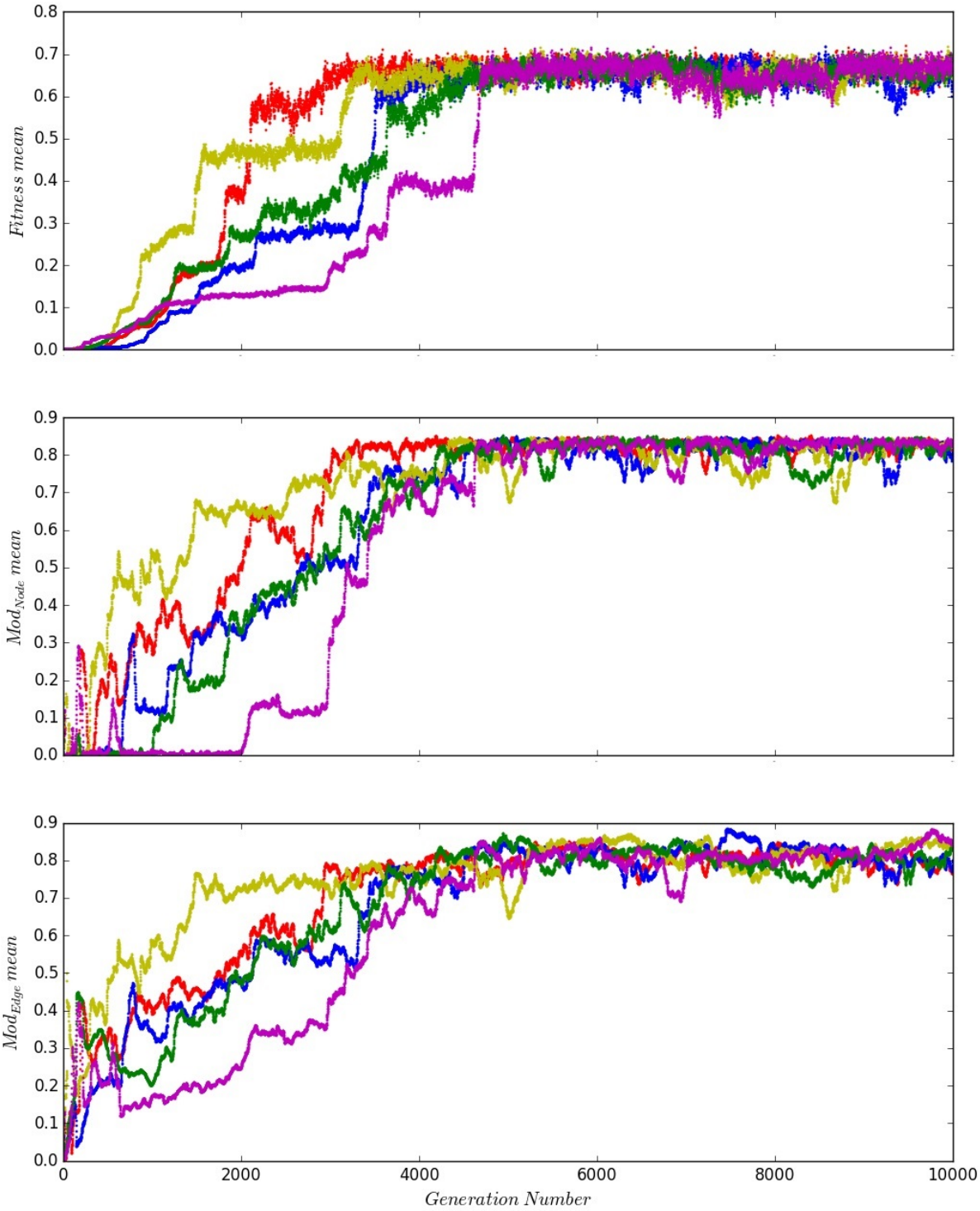

Figura 3.10: Fitness mean values, $\text{Mod}_{\boldsymbol{node}}$ mean values and $\text{Mod}_{\boldsymbol{edge}}$ mean values temporal evolution profile during 10000 generations. These mean values are calculated with the correspondent value of each one of the 1000 individuals in each generation. The input parameters of this essay were $\boldsymbol{prob_{duplic} = 0.005}$, $\boldsymbol{prob_{mut} = prob_{elim} = 0.001}$ and using the Environment Set **8OV**. Red dots: first replica. Blue dots: second replica. Yellow dots: third replica. Green dots: fourth replica. Purple dots: fifth replica.

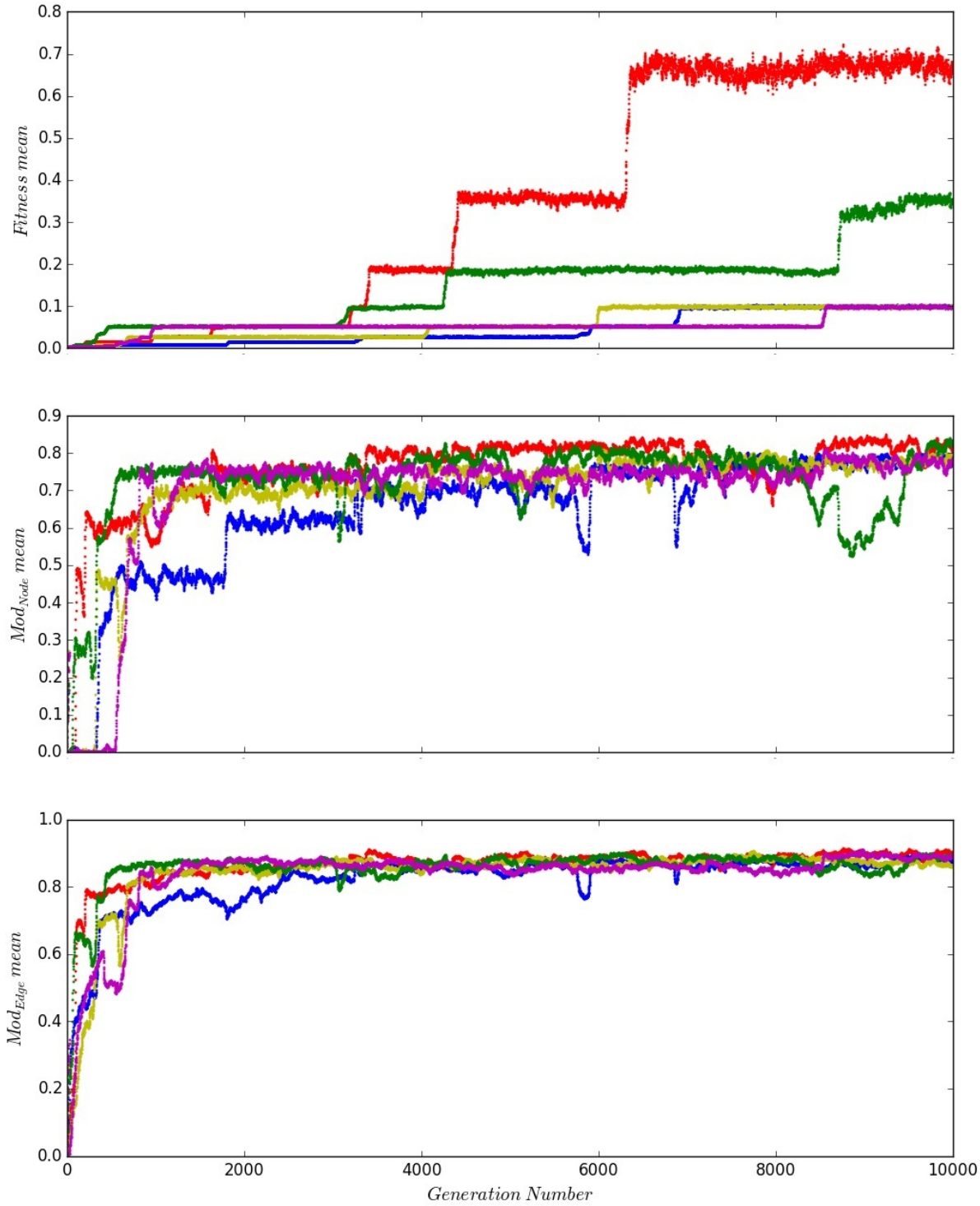

Figura 3.11: Fitness mean values, $\text{Mod}_{\boldsymbol{node}}$ mean values and $\text{Mod}_{\boldsymbol{edge}}$ mean values temporal evolution profile during 10000 generations. These mean values are calculated with the correspondent value of each one of the 1000 individuals in each generation. The input parameters of this essay were $\boldsymbol{prob_{duplic} = 0.005}$, $\boldsymbol{prob_{mut} = prob_{elim} = 0.001}$ and using the Environment Set **8NO**. Red dots: first replica. Blue dots: second replica. Yellow dots: third replica. Green dots: fourth replica. Purple dots: fifth replica.

It is possible to see a quantitative decoupling between the fitness and modularity evolution. In the simulations with 4 environments is easy to identify generations in which modularity significantly gets higher and it is accompanied by subtle fitness value increases. With 8 non-overlapped environments, a huge fitness variation is visible but that variation is not accompanied by the modularity values. In some replicas, it is possible to observe periods where the fitness mean values do not significantly change, but there is a negative variation for the modularity mean values. In the green replica (4th replica) it is possible to observe a positive variation in the fitness mean values and a negative variation in the modularity mean values. In this non-overlapped 8 environments simulations, it is also visible that modularity has a positive evolution, on the first generations, but mean values of fitness barely get any higher. This suggests that the way a network evolves, by edge mutation and node duplications and eliminations, can lead to modularity growths independently of associated fitness values.

3.3 High Fit Networks

A different way to approach impact of both the number of environments and overlap state in the signalling networks evolution is to select, within every single simulation that was made, with every parameter combination and every replica, the ones that completed their evolution with a high mean fitness value ($Fitness \geq 0.85$).

This analysis (**Figure 3.12**) shows that an increasing number of environments, makes higher fitness signalling networks evolution harder and that the environments overlap presence decreases the success rate.

Without overlapped environments, these high fit networks are very homogeneous in terms of really high modularity mean values, mean number of nodes and edges. As for the overlapped environments, it is possible to see two networks clusters, in terms of modularity, and a more sparse mean number of both nodes and edges(**Figure 3.13**). This suggests that is possible to obtain by natural selection networks with high values of fitness and relatively low modularity values.

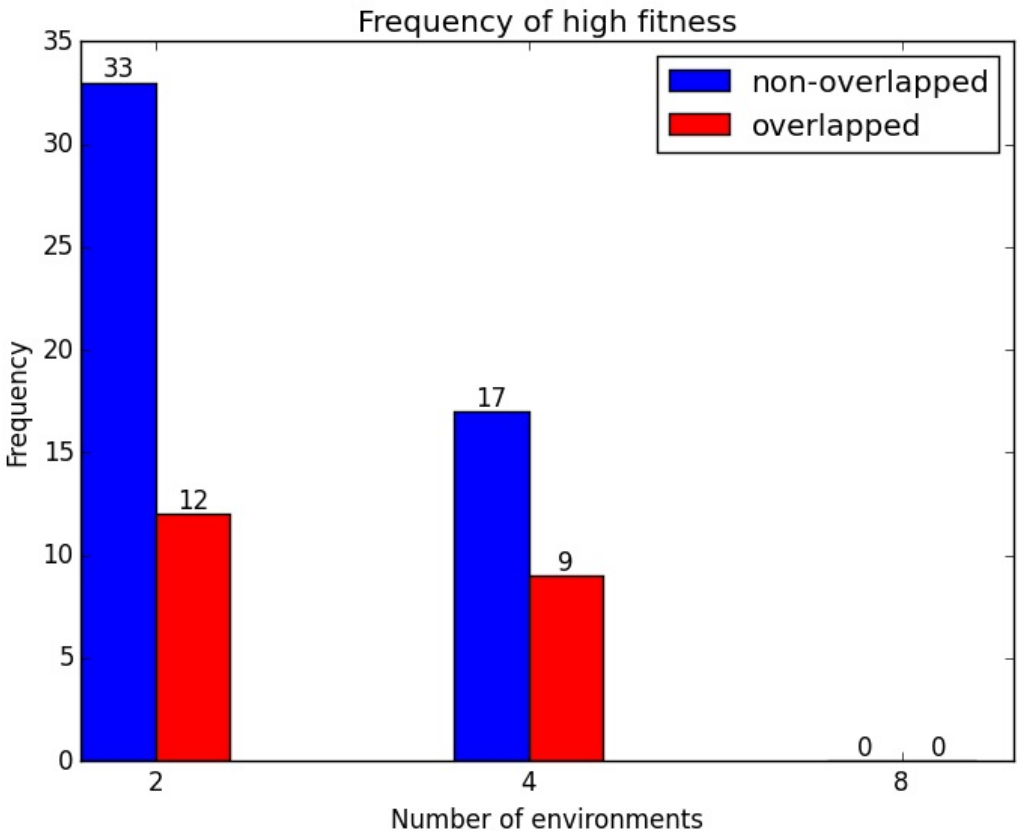

Figura 3.12: Histogram of the absolute frequency of the samples who showed higher values of fitness organized by number of environments. The **blue bars** represent the non-overlapped samples and the **red bars** represent the overlapped samples.

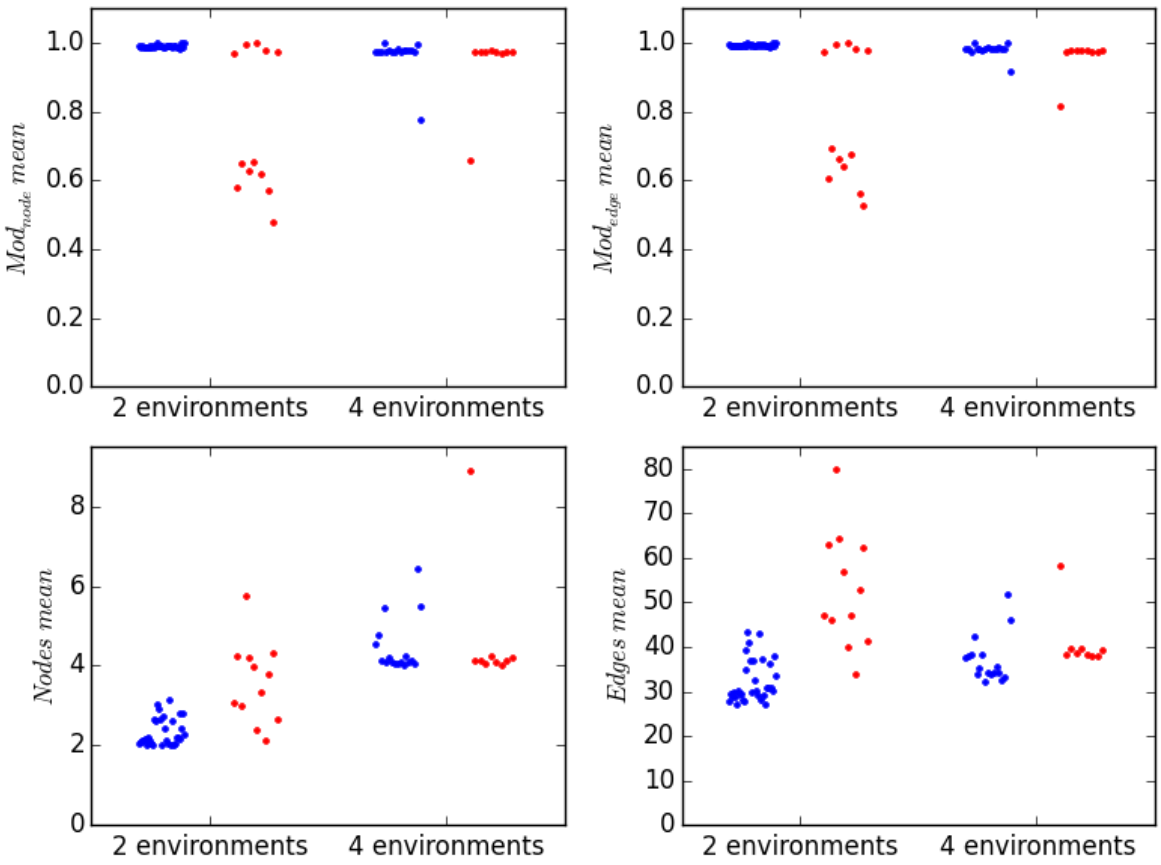

Figura 3.13: Visual representation of the samples populations with higher fitness values according to their modularity, edges and nodes. **Blue dots**: non-overlapped samples (**Tables 3.7 and 3.9**). **Red dots**: overlapped samples (**Tables 3.8 and 3.10**).

To evaluate if some mutational parameters had any impact on the clustering of modularity values, we tried, unsuccessfully, to find a pattern within the high fit organisms. This means that, through stochastic events, organisms can evolve their network's architecture to a non-modular one that grant them similar fitness to the ones that have modular networks. The mutational parameters used, as well as the set where the organisms are from and replica can be found through **Tables 3.7** to **3.10**,

Tabela 3.7: Mutational parameters that generated high fit samples for 2NO. Fitness, Mod_{node}, Mod_{edge}, Nodes and Edges are mean values of the 1000 individuals population at the 10000th generation

Set (replica nº)	$prob_{duplic}$	$prob_{mut}$	$prob_{elim}$	**Fitness**	Mod_{node}	Mod_{edge}	**Nodes**	**Edges**
001duplic005 (3)	0.005	0.001	0.001	0.9264	0.9829	0.9873	2.80	36.41
001all (2)	0.001	0.001	0.001	0.9370	0.9847	0.9888	2.05	27.98
001elim0001 (1)	0.001	0.001	0.0001	0.9432	0.9853	0.9895	2.15	28.94
001elim0001 (4)	0.001	0.001	0.0001	0.9373	0.9855	0.9899	2.19	30.08
005mut001 (1)	0.001	0.005	0.001	0.9297	0.9857	0.9900	2.63	37.30
001elim005 (4)	0.001	0.001	0.005	0.9344	0.9858	0.9902	2.09	29.40
005mut001 (2)	0.001	0.005	0.001	0.9183	0.9864	0.9897	2.64	34.78
001all (3)	0.001	0.001	0.001	0.9236	0.9867	0.9910	2.16	30.91
001elim005 (5)	0.001	0.001	0.005	0.9334	0.9868	0.9910	2.06	29.35
005mut001 (5)	0.001	0.005	0.001	0.9322	0.9869	0.9907	2.43	32.67
001duplic005 (1)	0.005	0.001	0.001	0.9361	0.9873	0.9908	2.72	37.07
001duplic005 (5)	0.005	0.001	0.001	0.9310	0.9876	0.9905	2.80	37.80
001all (4)	0.001	0.001	0.001	0.9492	0.9878	0.9913	2.22	30.73
001duplic0001 (3)	0.0001	0.001	0.001	0.9363	0.9880	0.9912	2.00	27.25
001elim005 (2)	0.001	0.001	0.005	0.9383	0.9880	0.9916	2.13	29.68
001all (5)	0.001	0.001	0.001	0.9334	0.9883	0.9918	2.21	30.91
001duplic005 (2)	0.005	0.001	0.001	0.9382	0.9886	0.9921	3.13	42.87
001elim0001 (2)	0.001	0.001	0.0001	0.9421	0.9887	0.9921	2.14	29.38
001elim0001 (5)	0.001	0.001	0.0001	0.9514	0.9887	0.9921	2.14	30.15
001duplic005 (4)	0.005	0.001	0.001	0.9312	0.9887	0.9926	2.63	39.44
005mut001 (4)	0.001	0.005	0.001	0.9360	0.9894	0.9927	2.67	36.96
001duplic0001 (1)	0.0001	0.001	0.001	0.9446	0.9895	0.9930	2.00	29.82
001duplic0001 (5)	0.0001	0.001	0.001	0.9421	0.9895	0.9929	2.00	29.08
001elim005 (1)	0.001	0.001	0.005	0.9355	0.9895	0.9922	2.06	27.13
001elim005 (3)	0.001	0.001	0.005	0.9467	0.9895	0.9925	2.13	28.69
005mut001 (3)	0.001	0.005	0.001	0.9255	0.9902	0.9931	2.90	41.07
001duplic0001 (2)	0.0001	0.001	0.001	0.9539	0.9905	0.9934	2.00	27.88
001duplic0001 (4)	0.0001	0.001	0.001	0.9504	0.9905	0.9931	2.00	28.06
001elim0001 (3)	0.001	0.001	0.0001	0.9427	0.9910	0.9937	2.06	27.91
001all (1)	0.001	0.001	0.001	0.9450	0.9923	0.9947	2.06	29.29
001mut0001 (2)	0.001	0.0001	0.001	0.9941	0.9987	0.9991	3.04	43.38
001mut0001 (3)	0.001	0.0001	0.001	0.9940	0.9988	0.9990	2.45	30.33
001mut0001 (1)	0.001	0.0001	0.001	0.9924	0.9995	0.9997	2.28	33.41

Tabela 3.8: Mutational parameters that generated high fit samples in the enviroment set (Env) 2OV. Fitness, Mod_{node}, Mod_{edge}, Nodes and Edges are mean values of the 1000 individuals population at the 10000th generation

Set (replica nº)	$prob_{duplic}$	$prob_{mut}$	$prob_{elim}$	**Fitness**	Mod_{node}	Mod_{edge}	**Nodes**	**Edges**
001elim0001 (2)	0.001	0.001	0.0001	0.9218	0.4775	0.5264	4.33	62.32
001all (4)	0.001	0.001	0.001	0.9371	0.5685	0.5626	3.78	53.00
005mut001 (3)	0.001	0.005	0.001	0.9196	0.5790	0.6048	4.24	62.87
001all(2)	0.001	0.001	0.001	0.9314	0.6183	0.6734	3.33	47.26
005mut001 (2)	0.001	0.005	0.001	0.9090	0.6271	0.6600	4.19	64.27
001duplic0001 (2)	0.0001	0.001	0.001	0.9336	0.6500	0.6946	2.99	45.98
005mut001 (4)	0.001	0.005	0.001	0.9110	0.6538	0.6417	3.96	56.87
001duplic005 (5)	0.005	0.001	0.001	0.9132	0.9668	0.9709	3.07	47.12
001elim0001 (4)	0.001	0.001	0.0001	0.9213	0.9731	0.9764	2.67	41.34
001elim005 (1)	0.001	0.001	0.005	0.9253	0.9772	0.9800	2.12	33.84
001mut0001 (3)	0.001	0.0001	0.001	0.9806	0.9943	0.9951	5.74	79.85
001mut0001 (1)	0.001	0.0001	0.001	0.9926	0.9992	0.9993	2.39	39.93

Tabela 3.9: Mutational parameters that generated high fit samples in the enviroment set (Env) 4NO. Fitness, Mod_{node}, Mod_{edge}, Nodes and Edges are mean values of the 1000 individuals population at the 10000th generation

Set (replica nº)	$prob_{duplic}$	$prob_{mut}$	$prob_{elim}$	**Fitness**	Mod_{nodes}	Mod_{edges}	**Nodes**	**Edges**
001elim0001 (3)	0.001	0.001	0.0001	0.8613	0.7740	0.9155	5.49	46.11
001elim0001 (2)	0.001	0.001	0.0001	0.8592	0.9713	0.9744	4.12	38.38
001elim0001 (1)	0.001	0.001	0.0001	0.8672	0.9728	0.9759	4.09	38.14
001all (4)	0.001	0.001	0.001	0.8692	0.9735	0.9809	4.09	33.88
001elim005 (1)	0.001	0.001	0.005	0.8610	0.9736	0.9803	4.07	32.28
001all (2)	0.001	0.001	0.001	0.8572	0.9742	0.9813	4.11	33.81
001duplic005 (4)	0.005	0.001	0.001	0.8562	0.9744	0.9809	4.76	38.04
001all (3)	0.001	0.001	0.001	0.8605	0.9746	0.9816	4.05	33.23
001duplic005 (3)	0.005	0.001	0.001	0.8568	0.9751	0.9819	4.56	37.79
001duplic0001 (4)	0.0001	0.001	0.001	0.8679	0.9758	0.9829	4.00	34.23
001all (1)	0.001	0.001	0.001	0.8676	0.9760	0.9830	4.26	35.56
001elim0001 (5)	0.001	0.001	0.0001	0.8729	0.9766	0.9837	4.20	35.21
001elim005 (3)	0.001	0.001	0.005	0.8537	0.9783	0.9836	4.12	32.70
001elim005 (5)	0.001	0.001	0.005	0.8643	0.9789	0.9851	4.10	34.31
001duplic0001 (5)	0.0001	0.001	0.001	0.8663	0.9798	0.9860	4.04	34.27
001mut0001 (1)	0.001	0.0001	0.001	0.9743	0.9963	0.9975	6.42	51.68
001mut0001 (5)	0.001	0.0001	0.001	0.9855	0.9981	0.9986	5.46	42.43

Tabela 3.10: Mutational parameters that generated high fit samples in the enviroment set (Env) 4OV. Fitness, Mod_{node}, Mod_{edge}, Nodes and Edges are mean values of the 1000 individuals population at the 10000th generation

Set (replica nº)	$\boldsymbol{prob_{duplic}}$	$\boldsymbol{prob_{mut}}$	$\boldsymbol{prob_{elim}}$	**Fitness**	$\boldsymbol{Mod_{node}}$	$\boldsymbol{Mod_{edge}}$	**Nodes**	**Edges**
001mut0001 (3)	0.001	0.0001	0.001	0.9536	0.657729973	0.8163	8.91	58.32
001duplic0001 (5)	0.0001	0.001	0.001	0.8551	0.9693	0.9729	4.00	37.99
001elim0001 (2)	0.001	0.001	0.0001	0.8592	0.9713	0.9744	4.12	38.38
001all (3)	0.001	0.001	0.001	0.8632	0.9714	0.9743	4.12	37.96
001elim005 (5)	0.001	0.001	0.005	0.8548	0.9720	0.9757	4.06	38.79
001elim0001 (1)	0.001	0.001	0.0001	0.8672	0.9728	0.9759	4.09	38.14
001all (2)	0.001	0.001	0.001	0.8599	0.9731	0.9769	4.14	39.50
001elim0001 (4)	0.001	0.001	0.0001	0.8716	0.9740	0.9770	4.19	39.27
001all (1)	0.001	0.001	0.001	0.8613	0.9754	0.9779	4.23	39.55

Capítulo 4

Discussion

Our results suggest that exposing an organism to different environments during its life time is sufficient to evolve modular regulatory networks. This observation is novel and complements the previous studies where modularity evolved due to cyclic changes of environment after a few generations [11, 10]. Our scenario of changing environments during lifetime is reasonable, as exemplified by organisms that live more than one year and have to adapt to seasonal climatic changes. Previous studies also requested common sub-problems across varying environments to efficiently select modular networks. Our results with non-overlapping environments show that modularity can evolve in the absence of common sub-problems if the environmental changes occur during the organism lifetime.

We also wanted to study the influence of the number of environments an organism has to deal with during his lifetime. The more environments they are exposed to, the more complex is the problem to solve. Therefore, with 8 different environments, a large number of generations is needed to obtain well adapted networks. Reversely, with less environments, well adapted populations quickly emerge after generating individuals with non-zero fitness. Although overall there is a positive correlation between fitness and modularity values, in some replicate simulations it is possible to detect a quantitative decoupling between both properties. This decoupling is more evident in more complex settings (more environments and overlapping environments). Additionally, among simulations reaching high fitness values, replicates with the same parameters and environment exposition produce equally fit populations but either a high or low modularity. This is consistent with the idea that modular networks are not necessarily more fit, and contradicts studies where a modular architecture is implied to directly affect an organism's fitness. [8].

Regarding the impact of the mutational parameters ($prob_{duplic}$, $prob_{mut}$ and $prob_{elim}$) on the network evolution, node elimination shows no significant contribution. This can be explained through the selection process, that in most cases is sufficient to reverse cases of over-duplication negative effects. To our knowledge, the impact of this parameter in signalling and gene expression regulatory networks modular architecture has not been addressed before.

The node duplication rate has a positive impact on both fitness and modularity of the organisms that are exposed to more complex problems, since there will be a need for more nodes to solve it. Many studies agree that gene duplication can lead adaptive advantages [5, 18, 19] and that it can contribute to a more modular architecture [5, 8]. However it has a negative impact on both fitness and modularity of organisms that are exposed to less complex problems. One possibility is that higher duplication creates an unnecessary excess of nodes. Our model, for simplification reasons, does not account for a fitness cost proportional to the number of connections or nodes. Introducing this kind of cost, which is biologically plausible [9], could diminish this negative impact of duplication.

The edge mutation rate has a negative impact on both fitness and modularity in most of the problems to which the organisms were presented. This is no surprise because, although it's a processes that it is needed to evolve the networks, it gets harder to retain partial solutions due to the fact that the network architecture will be changing very often.

The way we calculated modularity was different from previous studies [11, 20]. The reason we did it in a new way is due the fact that, according to our hypothesis, the number of modules that would arise from the simulations would be a number close to the number of environments to which the organisms were exposed during their life time [8, 10, 11]. Knowing the number of clusters and identifying cluster nodes and edges through their activation in each environment greatly simplifies modularity computation and simultaneously attributes a clear biological meaning to cluster membership.

The process of adding nodes to our networks was done by duplication previously existing nodesand copying the connections of the ancestor node. In a recent study, the duplication process is implemented in a different way: new nodes establish new connections randomly and connections from previously existing nodes are immutable [21]. We believe that every node must have a duplication probability each generation and every edge within the network must have mutation probability so it matches a more logic biological process [18, 19].

Our results are potentially influenced by simplifications in our model definition. First we used a boolean network model, where each node has only two states (on/off), second, all connections are activations and third, when a node has multiple inward edges, one active signal is sufficient for node activation (logical OR function). Other simplification that was made was the inability of inter-regulation between two nodes, e.g. if there is already an edge $N_0 \longrightarrow N_1$ in the graph, the creation of the edge $N_0 \longleftarrow N_1$ becomes impossible. Boolean networks have been successfully used to study signalling networks, capturing most essential dynamic properties [3], although we cannot exclude that quantitative effects may play some role in network evolution. Negative interactions and different logical functions to integrate multiple signals will be implemented in future versions of the model. Still, our results will be relevant to compare the effect of these more complex network properties.

Capítulo 5

Conclusions

In the past two decades, modularity has been a research topic and there has been a lot of discussion involving the origin of this network architecture. The main objective of this work was to prove that exposing organisms to different environments during their life time alone is a condition that can lead signalling and gene expression regulatory networks to achieve modularity. Our results support this hypothesis.

Our study also showed that the mutational parameters that drive the evolution of these networks have an impact on both fitness, modularity and overall organization. Higher node duplication rates are important to achieve higher fitness and modularity on more complex problems, since there will be needed a higher number of nodes to solve it. Node elimination rates do not seem to have any impact whatsoever. As for connection mutation rates, a slower mutation rate will improve the network's fitness and modularity, but it has to be a relatively slow, to allow the more fit and modular networks to retain the traits when they appear and this would not be possible with fast mutation.

Regarding the environmental changes impact on the network evolutionary process, the more complex (more and overlapping environments) the problem is, the later the network adapts to its surroundings. It was also visible, mainly in these complex scenarios, a decoupling between fitness and modularity evolution.

One interesting finding is that high fit networks are not necessarily modular. It is possible to achieve other architectures that are also fit, even with the same mutational parameters and environmental settings that were responsible for highly modular networks. This is an evidence that fitness and modularity are not necessarily correlated, but modularity can, indeed contribute to an adaptation to the surrounding environmental dynamics.

In the future, we want to make our computational model more complex to allow a more detailed study of the digital organisms. Adding a connection cost to match the new models of evolution may improve the accuracy of our data when the organisms are exposed to less complex problems. We also want to be able to directly control the connection mutation rate, separating this parameter in two: one related to the addition of new connections and other related the the elimination of previously existing connections. In

order to improve the biological accuracy of our model it is also important to allow different types of regulation between genes to occur since currently it only accounts for positive and boolean regulation. It may be also interesting to add a cross-over function to the digital network to simulate more complex organisms.

Bibliografia

[1] Willi Gottstein, Stefan Müller, Hanspeter Herzel, and Ralf Steuer. Elucidating the adaptation and temporal coordination of metabolic pathways using in-silico evolution. *BioSystems*, 117(1):68–76, 2014.

[2] Arend Hintze and Christoph Adami. Evolution of complex modular biological networks. *PLoS Computational Biology*, 4(2), 2008.

[3] István Albert, Juilee Thakar, Song Li, Ranran Zhang, and Réka Albert. Boolean network simulations for life scientists. *Source code for biology and medicine*, 3:16, 2008.

[4] Guillermo Rodrigo, Javier Carrera, and Santiago F. Elena. Network design meets in silico evolutionary biology. *Biochimie*, 92(7):746–752, 2010.

[5] Wynand Winterbach, Piet Van Mieghem, Marcel Reinders, Huijuan Wang, and Dick de Ridder. Topology of molecular interaction networks. *BMC systems biology*, 7(1):90, 2013.

[6] Wenzhe Ma, Ala Trusina, Hana El-Samad, Wendell a. Lim, and Chao Tang. Defining Network Topologies that Can Achieve Biochemical Adaptation. *Cell*, 138(4):760–773, 2009.

[7] Wenzhe Ma, Luhua Lai, Qi Ouyang, and Chao Tang. Robustness and modular design of the Drosophila segment polarity network. *Molecular systems biology*, 2:70, 2006.

[8] Günter P Wagner, Mihaela Pavlicev, and James M Cheverud. The road to modularity. *Nature reviews. Genetics*, 8(12):921–931, 2007.

[9] Jeff Clune, Jean-baptiste Mouret, and Hod Lipson. The evolutionary origins of modularity. *Proceedings. Biological sciences / The Royal Society*, 280(1755):20122863, 2013.

[10] Nadav Kashtan, Elad Noor, and Uri Alon. Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13711–13716, 2007.

[11] Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–13778, 2005.

[12] Yusuke Ikemoto and Kosuke Sekiyama. Modular network evolution under selection for robustness to noise. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 89(4):1–11, 2014.

[13] Tien-Dzung Tran and Yung-Keun Kwon. The relationship between modularity and robustness in signalling networks. *Journal of the Royal Society, Interface / the Royal Society*, 10(88):20130771, 2013.

[14] Claus O. Wilke and Christoph Adami. The biology of digital organisms. *Trends in Ecology and Evolution*, 17(11):528–532, 2002.

[15] Bradly Alicea and Richard Gordon. Toy models for macroevolutionary patterns and trends. *BioSystems*, 122:25–37, 2014.

[16] Mihaela E Sardiu, Yong Cai, Jingji Jin, Selene K Swanson, Ronald C Conaway, Joan W Conaway, Laurence Florens, and Michael P Washburn. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5):1454–1459, 2008.

[17] Marti J. Anderson and Pierre Legendre. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62(3):271–303, 1999.

[18] R Gordon. Evolution escapes rugged fitness landscapes by gene or genome doubling: the blessing of higher dimensionality. *Computers & chemistry*, 18(3):325–31, 1994.

[19] P. Dwight Kuo, Wolfgang Banzhaf, and André Leier. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems*, 85(3):177–200, 2006.

[20] John J Welch and David Waxman. Modularity and the cost of complexity. *Evolution; international journal of organic evolution*, 57(8):1723–1734, 2003.

[21] Gourab Ghoshal, Liping Chi, and Albert-Laszlo Barabasi. Uncovering the role of elementary processes in network evolution. *Scientific reports*, 3:2920, 2013.