Population Exploration on Genotype Networks in Genetic Programming

Ting Hu[1], Wolfgang Banzhaf[2], and Jason H. Moore[1]

[1] Computational Genetics Laboratory, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03756, USA
{ting.hu,jason.h.moore}@dartmouth.edu
[2] Department of Computer Science, Memorial University, St. John's, NL, A1B 3X5, Canada
banzhaf@mun.ca

Abstract. Redundant genotype-to-phenotype mappings are pervasive in evolutionary computation. Such redundancy allows populations to expand in neutral genotypic regions where mutations to a genotype do not alter the phenotypic outcome. Genotype networks have been proposed as a useful framework to characterize the distribution of neutrality among genotypes and phenotypes. In this study, we examine a simple Genetic Programming model that has a finite and compact genotype space by characterizing its genotype networks. We study the topology of individual genotype networks underlying unique phenotypes, investigate the genotypic properties as vertices in genotype networks, and discuss the correlation of these network properties with robustness and evolvability. Using GP simulations of a population, we demonstrate how an evolutionary population diffuses on genotype networks.

1 Introduction

A remarkable feature of natural evolutionary systems is how they maintain resilience to constant intrinsic and environmental perturbations while remaining adaptive in the face of survival challenges. Robustness [1, 2] and evolvability [3–5] have been discussed as closely related but somewhat contradictory properties in this context. Essentially, both properties reflect how evolutionary systems respond to changes. Robustness enables them to remain intact in the face of deleterious changes, whereas evolvability allows them to innovate to better fit the survival pressures of the environment. Redundancy is a crucial mechanism contributing to both robustness and evolvability. A redundant mapping from multiple genotypes to a phenotype allows genetic variants to expand in neutral mutational spaces. These neutral spaces are genotypic regions in which mutations do not change the phenotype or fitness. Neutral genetic variations by mutations possess the potential for creating novel phenotypes [6]. They serve as a quantitive staging ground for long-term adaptation and innovation. Such neutrality provides a buffer against deleterious mutational perturbations, and augments evolvability by accumulating genetic variations that might be non-neutral under changes of the environmental context [7–10].

T. Bartz-Beielstein et al. (Eds.): PPSN XIII 2014, LNCS 8672, pp. 424–433, 2014.

Genotype networks, a.k.a. neutral networks, have been proposed as a useful framework for studying neutrality [11–13]. In such networks, genotypes are represented as vertices, and reversible mutational connections, as in common evolutionary systems, are represented as undirected edges between pairs of genotypes. One genotype network is comprised of all genotypes that encode for the same phenotype. Therefore, within a genotype network, edges denote only neutral point mutations. Different genotype networks, i.e. phenotypes, can also be connected through non-neutral point mutations between genotypes that are phenotypically distinguished. Genotype networks provide a global view of how neutrality is distributed among various phenotypes, and hence become a very useful framework to investigate how redundancy contributes to robustness and evolvability. On one hand, studies have shown that evolutionary search really benefits from expanding neutral regions [14, 15]. On the other hand, some evolutionary systems are found to be constrained by the abundance of neutral mutational variants [16].

A redundant mapping from genotype to phenotype is also pervasive in many Evolutionary Computation (EC) systems, especially in Genetic Programming (GP), where multiple genotypes encode identical phenotypes [17–19]. A single point mutation to a genotype is defined as neutral if it does not alter the phenotype or fitness. Such neutrality is largely contributed by the considerable amount of non-coding regions in GP. Departing from early recognition of these non-coding regions as disadvantageous, later extensive investigations and discussions have been conducted on how to characterize and utilize neutrality in GP [20–22]. The notion of genotype networks has also been adopted in many GP neutrality studies [23, 24]. However, most studies characterizing genotype networks are constrained by the infeasibility of enumerating genotypes due to the infinite genotypic space of common GP systems. In a recent study, a quantitive characterization of mutational robustness and evolvability was performed using a simple Linear GP model, where the entire genotype and phenotype spaces are finite and enumerable [25]. It is reported that robustness and evolvability are correlated in a different way at the genotypic, phenotypic, and fitness levels.

In this study, we adopt the same Linear GP model as in a quantitative study on evolvability and robustness [25] to take advantage of its genotype space being amenable to exhaustive enumeration. We characterize topological properties of individual genotype networks and take a close look at vertex importance of genotypes in the networks and how it correlates with robustness and evolvability. Furthermore, using GP simulations, we investigate how an evolutionary population diffuses on genotype networks and how those movements on the genotype networks are reflected in fitness improvements.

2 Methods

2.1 Problem Instance

We consider a simple Linear GP system on a Boolean search problem as in a previous study [25]. In the LGP representation, an individual (or computer program)

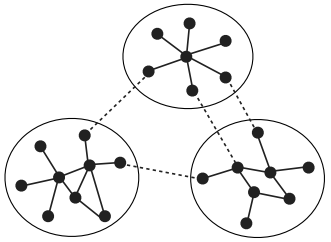

Fig. 1. Schematic diagram of a subset of genotype networks. Each vertex represents a genotype and genotypes encoded to the same phenotype form one genotype network. An edge links two vertices if the two genotypes can be transformed from one to another through a single point mutation. Single point mutations can also connect genotypes from different phenotypes, shown in dashed lines.

consists of a set of L instructions, which are structurally similar to those found in register machine languages. Each instruction has an operator, a set of operands, and a return value. In our study, each instruction consists of an operator drawn from the Boolean function set {AND, OR, NAND, NOR}, two Boolean operands, and one Boolean return value. The inputs, operands, and return values are stored in registers with varying read/write permissions. Specifically, $\mathtt{R}_0$ and $\mathtt{R}_1$ are calculation registers that can be read and written, whereas $\mathtt{R}_2$ and $\mathtt{R}_3$ are input registers that are read-only. Thus, a calculation register can serve in an instruction as an operand or a return, but an input register can only be used as an operand. An example program with $L = 3$ is given below.

$$\begin{aligned}
&\mathtt{R}_1 = \mathtt{R}_2 \ \ \mathtt{AND} \ \ \mathtt{R}_3 \\
&\mathtt{R}_0 = \mathtt{R}_2 \ \ \ \mathtt{OR} \ \ \ \mathtt{R}_1 \\
&\mathtt{R}_0 = \mathtt{R}_3 \ \mathtt{NAND} \ \mathtt{R}_0
\end{aligned}$$

These instructions are executed sequentially from top to bottom. Prior to program execution, the values of $\mathtt{R}_0$ and $\mathtt{R}_1$ are initialized to FALSE. Registers $\mathtt{R}_2$ and $\mathtt{R}_3$ read two Boolean input values. After program execution, the final value in $\mathtt{R}_0$ is returned as output.

2.2 Genotype, Phenotype, and Genotype Networks

We consider each unique LGP program as a *genotype* and the binary Boolean function $f : \mathbf{B}^2 \to \mathbf{B}$, where $\mathbf{B} = \{\mathtt{TRUE}, \mathtt{FALSE}\}$, represented by the program as its *phenotype*. We set two calculation registers, two input registers and four operators, which means there are $2 \times 4 \times 4 \times 4 = 2^7$ possible instructions and thus 2^{21} possible programs of length $L = 3$. These 2^{21} programs define the finite genotype space mapping to the 16 possible binary Boolean functions $f : \mathbf{B}^2 \to \mathbf{B}$ as phenotypes.

Genotypes transform from one to another through point mutations. These mutational connections can be well modeled by networks. The framework of

genotype networks has been proposed to study how mutational connections are distributed among genotypes underlying various phenotypes [11–13]. A *genotype network* is comprised of all the genotypes, as vertices, that represent the same phenotype. An edge connects a pair of genotypes if they can be transferred from each other through a single point mutation (see Fig. 1).

Different phenotypes can have varying genotype network properties, and investigating these network properties provides insights into how an evolutionary population explores the genotype space by expanding in genotype networks. We take advantage of the simple yet representative LGP system to fully characterize the entire genotype space by enumerating all genotypes and constructing all 16 genotype networks. Then, for each genotype network, we look at their network properties including network size, i.e. the total number of vertices, network degree distribution and vertex closeness centrality.

The degree of a vertex in a network is the number of its connected neighbors. In the framework of genotype networks, vertex degree reflects how robust a genotype is when subject to point mutations. High degree vertices are genotypes that are more likely to maintain their phenotypes under point mutations. Vertex degree distribution describes the global connectivity of a network. At the vertex level, centrality measures the importance of a vertex in the network. There are a number of centrality measures that capture the individual contribution of vertices to a network. In the current study, we look at the *closeness centrality*, denoted as $\frac{1}{\sum_{j \neq i} d_{ij}}$ of a vertex i, where d_{ij} is the distance, i.e. the shortest path, between vertices i and j [26, 27]. Closeness centrality describes how easily a given vertex can reach all other vertices. A higher closeness centrality indicates a more central position of a vertex in the network. Beyond mutational connections within genotype network, genotypes can mutate into different phenotypes through non-neutral single point mutations. The *evolvability* of a genotype is defined as the number of unique phenotypes that it can reach through single point mutations [13]. This definition is intuitive in that if a genotype is adjacent to genotypes from many other different phenotypes, it is considered more evolvable.

2.3 Population Evolution

Population evolution is simulated to investigate how a population diffuses on genotype networks. The initial population includes $|P|$ randomly chosen genotypes from one given phenotype. Then for each generational iteration, a number of individuals are subject to single point mutations, according to a mutation rate r, and both $|P|$ parents and $r \times |P|$ offspring are competing in a tournament selection to form the next generation of $|P|$ individuals. We set one particular phenotype as the target and let the population evolve towards it. The evolution process is terminated once the entire population converges to the target.

A single point mutation can apply to any locus of a genome, including the return register, one of the two operand registers, or the operation function. The fitness value of a genotype is calculated based on the mutational potential from its phenotype to the target phenotype. Specifically, let v_{ij} denote the total

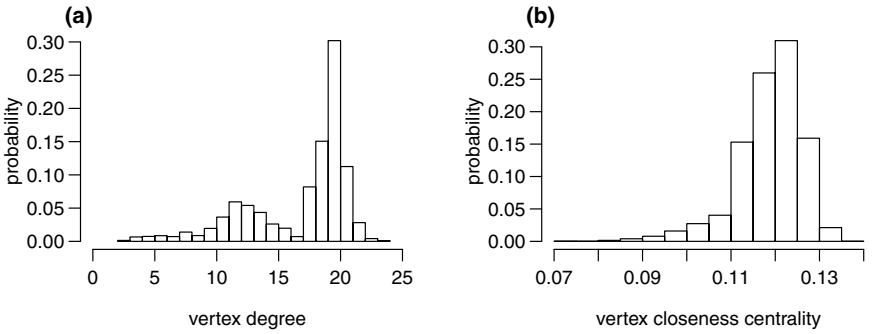

Fig. 2. Genotype network properties of phenotype `NAND`. (a) Distribution of vertex degree. (b) Distribution of vertex closeness centrality.

number of possible single point mutations that can transform genotypes from phenotype i to phenotype j. The fitness of genotypes from phenotype k with regard to the target phenotype t is defined as $f_t(k) = \frac{v_{kt}}{\sum_{j \neq k} v_{kj}}$. This fitness calculation is defined following the intuition that a phenotype with a higher mutational potential towards the target is rewarded with a higher fitness value.

3 Results and Discussion

3.1 Properties of Genotype Networks

For our particular LGP system, the distribution of genotypes among different phenotypes is highly heterogeneous. The size of genotype networks ranges from a minimum of 64 genotypes (for phenotypes `EQUAL` and `XOR`) to a maximum of 617,024 genotypes (for `FALSE`), occupying between $\ll 0.1\%$ and 29.4% of the entire genotype space, respectively. The mutational connections among phenotypes are also unevenly distributed. Out of the total 16 genotype networks, 10 are mutationally accessible from all other genotype networks, 4 are adjacent to 14 other phenotypes, and the two smallest genotype networks (`EQUAL` and `XOR`) have 13 phenotype neighbors. Moreover, these two smallest genotype networks are comprised of 64 individual islands, i.e. all 64 genotypes mutate away from their phenotype with any single point mutations, whereas the other 14 genotype networks contain single connected components.

Due to the symmetry of Boolean functions, some genotype networks share the same topological properties, e.g. phenotypes `x >= y` and `x <= y`. Interestingly, all genotype networks, excluding `EQUAL` and `XOR` that have all genotypes as isolated vertices, share the bi-modal vertex degree distribution. Fig. 2(a) shows the degree distribution of the representative genotype network `NAND`. This degree distribution suggests that the genotype networks are comprised of a dense core of highly connected genotypes, as well as a cluster of genotypes towards the periphery. The vertex closeness centrality has a uni-modal distribution (Fig. 2(b)),

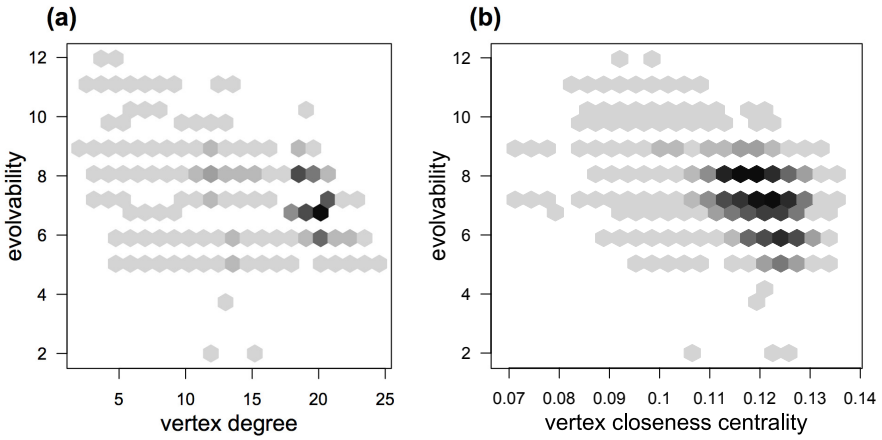

Fig. 3. The correlations of genotypic evolvability and (a) vertex degree and (b) vertex closeness centrality of the `NAND` genotype network. Grayscale of hexagons indicates the density of value intervals.

suggesting that most vertices are about equally accessible from other vertices in the network. The vertex degree and closeness centrality are also positively correlated (Spearman's rank correlation $\rho = 0.5417$, $p < 2 \times 10^{-16}$).

Recall that the evolvability of a given genotype is measured as the number of accessible unique phenotypes through single point mutations. We then look into how genotypic evolvability correlates with the degree and closeness centrality of a genotype in the network. Fig. 3 shows evolvability as a function of (a) vertex degree and (b) vertex closeness centrality. It can be observed that both the vertex degree and closeness centrality are negatively correlated with evolvability (Spearman's rank correlations $\rho = -0.3379$, $p < 2 \times 10^{-16}$ and $\rho = -0.3865$, $p < 2 \times 10^{-16}$, respectively). This suggests that the dense center cores of genotype networks have less access to other unique phenotypes, i.e. are less evolvable, than the genotypes at the periphery.

3.2 Population Diffusion on Genotype Networks

After quantifying the static properties of genotype networks, we now use population evolution to investigate how a population diffuses on genotype networks. We set one of the least representative phenotypes, `x == y`, as the evolution target to allow evolution to proceed for a longer time. $|P| = 500$ individuals of a given starting phenotype are randomly sampled as the initial population. We set mutation rate $r = 0.1$ and use a tournament selection of size two. For each starting phenotype configuration we collect 1,000 independent runs, and for each run the required number of generations that a population converges to the target phenotype is recorded.

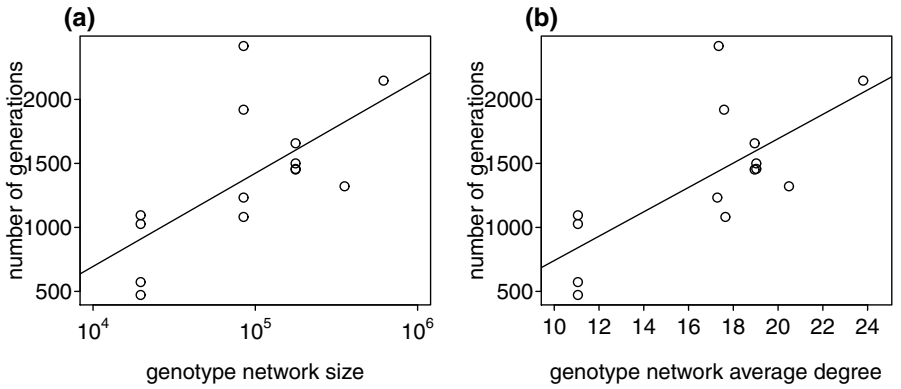

Fig. 4. The required evolution time as a function of the starting phenotype's genotype network properties. (a) The number of required generations increases as the genotype network becomes larger. (b) Evolution also requires a longer time if the starting phenotype's average vertex degree is larger. The lines provide a visual guide of their correlations.

Fig. 4 shows the correlations of the required evolution time and the starting phenotype's properties. A population needs a longer time to reach and converge to the target if it starts from a larger genotype network (Fig. 4(a), Spearman's rank correlation $\rho = 0.6640$, $p = 0.0096$). This positive correlation also exists between the evolution time and the starting phenotype's average vertex degree (Fig. 4(b), Spearman's rank correlation $\rho = 0.6326$, $p = 0.0152$). This suggests that it takes a population longer to evolve if it starts from a larger and more connected genotype network. This finding contradicts Wagner's RNA results where larger phenotypes, i.e. more robust phenotypes, are more evolvable [13], but agrees with Cowperthwaite's argument that the abundance of genotype networks constrains evolution [16]. We would like to point out that their correlation crucially depends on the properties of an evolutionary system, specifically how the genotype networks are adjacent to each other globally and where the target phenotype is located. For our LGP system, the target phenotype can be accessed from 13 other phenotypes, such that there are many possible paths to find the target. Therefore, evolution is expected to take a longer time moving out of large genotype networks and exploring novel phenotypes.

Last we take a close look at how a population diffuses on genotype networks as evolution proceeds. Fig. 5 shows the average vertex degree of a population changes as a function of generation in a typical run. The population starts with the average vertex degree of the starting genotype network `NAND`. Individuals then quickly move towards the periphery of the networks (generation 1 to 10). During the subsequent search, the population visits many other phenotypes, but leaves without going into their center cores (generation 11 to 160). The population reaches the first genotype of the target phenotype at generation 161, and quickly converges to the target in the next 33 generations. Also note that,

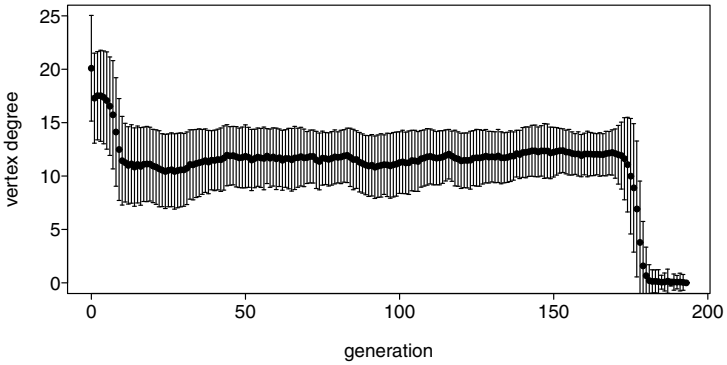

Fig. 5. The change of average vertex degree of an evolving population as evolution proceeds in a typical run of starting phenotype `NAND`. Points are population mean and error bars are standard deviations.

data not shown here, the change of vertex closeness centrality follows the same trend as the vertex degree since they are positively correlated.

4 Concluding Remarks

Here we have used a simple yet representative LGP system to fully characterize all individual mutational genotype networks by exhaustively enumerating the entire genotype and phenotype spaces. The 16 unique phenotypes are represented by 16 genotype networks that possess both shared and distinguishing properties. The two smallest genotype networks are comprised of isolated individual vertices, whereas all other networks contain single connected components. The connected genotype networks share similar bi-modal degree distributions, which indicate that the networks are comprised of a dense core and a well-connected periphery. In such genotype networks, vertices with high degrees are more likely located in the center of connecting all other vertices. However, these high-degree and high-centrality genotypes are less evolvable towards novel phenotypes.

By simulating population evolution, we find that a population requires more time to find a target if it starts from a larger genotype network. This observation conforms well to the static characterization of genotype properties in networks. We would like to point out that how the abundance of mutational variants contribute to evolvability crucially depends on the distribution of neutrality among various phenotypes and where the target phenotype is located. Our simulation also shows how an evolutionary population diffuses on genotype networks. It moves from the center of a network towards the periphery as the evolutionary search proceeds, accompanied by fitness improvements, and stays on the periphery of genotype networks visited until the target phenotype is reached.

The findings of this study provide insights on how neutrality is distributed in a typical LGP system. We conjecture that genotype networks could be shaped

very differently in other GP systems, however our current observations capture many general properties of GP, and might even be applicable to other EC systems. Specifically, the distribution of neutrality is very heterogenous among various phenotypes. Some genotype networks, i.e. phenotypes, could be orders of magnitude larger than others. Moreover, the mutational connections among phenotypes are biased, where a phenotype has more potential to mutate to particular phenotypes and is less likely to mutate to or is even disconnected from some phenotypes. The success of an innovative evolutionary search crucially depends on locating the target phenotype, i.e. whether it is accessible from many other phenotypes, and on finding an efficient mutational path towards it.

In future studies, we expect to use our methodology in other GP- or EC-systems and test if our observations and conjectures hold for a wider range of applications. It would be helpful to look into how a particular EC representation correlates with genotype network properties, such that we can gain a better understanding of how a representation influences evolutionary search and how we could improve the performance of an evolutionary algorithm by designing more appropriate representations.

Acknowledgments. This work was supported by National Institute of Health (USA) grants R01-LM009012, R01-LM010098, R01-AI59694, P20-GM103506, and P20-GM103534. W.B. acknowledges support from NSERC Discovery Grants, under RGPIN 283304-2012.

References

1. Lenski, R.E., Barrick, J.E., Ofria, C.: Balancing robustness and evolvability. PLoS Biology 4(12), e428 (2006)
2. van Nimwegen, E., Crutchfield, J.P., Huynen, M.A.: Neutral evolution of mutational robustness. Proceedings of the National Academy of Sciences 96(17), 9716–9720 (1999)
3. Kirschner, M., Gerhart, J.: Evolvability. Proceedings of the National Academy of Sciences 95, 8420–8427 (1998)
4. Pigliucci, M.: Is evolvability evolvable? Nature Review Genetics 9, 75–82 (2008)
5. Wagner, A.: Robustness, evolvability, and neutrality. Federation of European Biochemical Societies Letters 579(8), 1772–1778 (2005)
6. Masel, J., Trotter, M.V.: Robustness and evolvability. Trends in Genetics 26, 406–414 (2010)
7. Draghi, J.A., Parsons, T.L., Wagner, G.P., Plotkin, J.B.: Mutational robustness can facilitate adaptation. Nature 463, 353–355 (2010)
8. Landry, C.R., Lemos, B., Rifkin, S.A., Dickinson, W.J., Hartl, D.L.: Genetic properties influcing the evolvability of gene expression. Science 317, 118–121 (2007)
9. McBride, R.C., Ogbunugafor, C.B., Turner, P.E.: Robustness promotes evolvability of thermotolerance in an RNA virus. BMC Evolutionary Biology 8, 231 (2008)
10. de Visser, J.A.G.M., Hermission, J., Wagner, G.P., Meyers, L.A., Bagheri-Chaichian, H., et al.: Evolution and detection of genetic robustness. Evolution 57(9), 1959–1972 (2003)

11. Reidys, C., Stadler, P.F., Schuster, P.: Generic properties of combinatory maps: neutral networks of RNA secondary structures. Bulletin of Mathematical Biology 59(2), 339–397 (1997)
12. Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L.: From sequences to shapes and back: A case study in RNA secondary structures. Proceedings of The Royal Society B 255, 279–284 (1994)
13. Wagner, A.: Robustness and evolvability: A paradox resolved. Proceedings of The Royal Society B 275(1630), 91–100 (2008)
14. Ciliberti, S., Martin, O.C., Wagner, A.: Innovation and robustness in complex regulatory gene networks. Proceedings of the National Academy of Sciences 104(34), 13591–13596 (2007)
15. Wilke, C.O.: Adaptive evolution on neutral networks. Bulletin of Mathematical Biology 63, 715–730 (2001)
16. Cowperthwaite, M.C., Economo, E.P., Harcombe, W.R., Miller, E.L., Meyers, L.A.: The ascent of the abundant: How mutational networks constrain evolution. PLoS Computational Biology 4(7), e1000110 (2008)
17. Banzhaf, W.: Genotype-phenotype mapping and neutral variation - a case study in genetic programming. In: Davidor, Y., Schwefel, H.P., Manner, R. (eds.) PPSN 1994. LNCS, vol. 866, pp. 322–332. Springer, Heidelberg (1994)
18. Rothlauf, F., Goldberg, D.E.: Redundant representations in evolutionary computation. Evolutionary Computation 11(4), 381–415 (2003)
19. Hu, T., Banzhaf, W.: Evolvability and speed of evolutionary algorithms in light of recent developments in biology. Journal of Artificial Evolution and Applications 568375 (2010)
20. Galvan-Lopez, E., Poli, R.: An empirical investigation of how and why neutrality affects evolutionary search. In: Cattolico, M. (ed.) Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1149–1156 (2006)
21. Hu, T., Banzhaf, W.: Neutrality and variability: Two sides of evolvability in linear genetic programming. In: Rothlauf, F. (ed.) Proceedings of the Genetic and Evolutionary Computation Conference, pp. 963–970 (2009)
22. Soule, T.: Resilient individuals improve evolutionary search. Artificial Life 12, 17–34 (2006)
23. Banzhaf, W., Leier, A.: Evolution on neutral networks in genetic programming. In: Yu, T., Riolo, R., Worzel, B. (eds.) Genetic Programming Theory and Practice III, pp. 207–221. Springer (2006)
24. Ebner, M., Shackleton, M., Shipman, R.: How neutral networks influence evolvability. Complexity 7(2), 19–33 (2002)
25. Hu, T., Payne, J.L., Banzhaf, W., Moore, J.H.: Evolutionary dynamics on multiple scales: A quantitative analysis of the interplay between genotype, phenotype, and fitness in linear genetic programming. Genetic Programming and Evolvable Machines 13, 305–337 (2012)
26. Bavelas, A.: Communication patterns in task-oriented groups. Journal of the Acoustical Society of America 22, 725–730 (1950)
27. Sabidussi, G.: The centrality index of a graph. Psychometrika 31(4), 581–603 (1966)