Nonsynonymous to Synonymous Substitution Ratio $k_{\mathrm{a}}/k_{\mathrm{s}}$: Measurement for Rate of Evolution in Evolutionary Computation

Ting Hu and Wolfgang Banzhaf

Department of Computer Science, Memorial University of Newfoundland, Canada
{tingh, banzhaf}@cs.mun.ca

Abstract. Measuring fitness progression using numeric quantization in an Evolutionary Computation (EC) system may not be sufficient to capture the rate of evolution precisely. In this paper, we define the rate of evolution R_{e} in an EC system based on the rate of efficient genetic variations being accepted by the EC population. This definition is motivated by the measurement of "amino acid to synonymous substitution ratio" $k_{\mathrm{a}}/k_{\mathrm{s}}$ in biology, which has been widely accepted to measure the rate of gene sequence evolution. Experimental applications to investigate the effects of four major configuration parameters on our rate of evolution measurement show that R_{e} well reflects how evolution proceeds underneath fitness development and provides some insights into the effectiveness of EC parameters in evolution acceleration.

1 Introduction

Evolutionary computation is a method that simulates natural evolution to search for solutions to optimization problems. This field has seen significant progress in the past decades. Improving the evolutionary capabilities of an evolutionary system has attracted substantial attention recently [1], particularly enabling an evolutionary computation system to generate *evolvable* adaptation to its environments. Various evolution rates accompany diverse evolutionary capabilities. Measuring the rate of evolution can help to quantify evolutionary capabilities, and thus can be used to accelerate evolution through designing better computation models. At the time of writing, the rate of evolution has not yet seen a formal definition in the literature other than measuring fitness progression over generations. At first glance, a definition reflecting how fast an evolutionary population is improving its fitness may seem sufficient. However, considered as the capabilities to generate adaptation, evolutionary capabilities cannot be determined by how good population fitness is per se, but should be regarded as a "second-order" effect of fitness improvements. Therefore, we believe that the rate of evolution should be better defined by looking beyond fitness and should be measured by the rate of genetic variations being generated and accepted.

Some methods to quantify evolutionary capabilities have been proposed in the literature. Bedau and Packard [3] proposed a method to identify the capabilities of creating adaptation during evolutionary processes. It is based on calculating *evolutionary activity* statistics of components in an evolutionary system. Their comparison between artificial and natural evolutionary systems by studying the evolutionary activities showed that the "long-term" trend of generating adaptation is missing in artificial systems, i.e., the capability of generating evolvable adaptation is not as strong in artificial evolutionary systems.

Biologists use the $k_\mathrm{a}/k_\mathrm{s}$ ratio in molecular evolution to measure the evolution rate of gene sequences [8, 9]. Such a measurement compares two homologous protein-coding gene sequences from two related species. The $k_\mathrm{a}/k_\mathrm{s}$ ratio resulting from measuring the number of nonsynonymous (amino acid) substitutions per nonsynonymous site (k_a) to the number of synonymous substitutions per synonymous site (k_s) characterizes the rate of evolution between these two sequences. Here, *substitutions* only include those observable genetic changes having been accepted into the gene sequences. Since k_s measures neutral evolution (without involving functional improvements under selection pressure), the $k_\mathrm{a}/k_\mathrm{s}$ ratio reflects the rate of *adaptive* evolution against the *background* rate of evolution. This measurement has been widely applied in the analysis of adaptive molecular evolution, and is regarded as a general method of measuring the rate of sequence evolution in biology.

In this paper, we introduce the measurement of this $k_\mathrm{a}/k_\mathrm{s}$ ratio to EC. We utilize a Genetic Programming system to implement measuring the rate of evolution. Specifically, the rate of evolution in a GP system and the measurement of this rate are defined here. Comparative experiments on varying parameters including tournament selection size, population size, mutation rate, and crossover rate show the effectiveness of this approach. It is able to capture the rate of generating adaptive variations, which cannot be well observed in fitness development. We conclude this paper with a brief discussion on some future research.

2 The $k_\mathrm{a}/k_\mathrm{s}$ Ratio in Biology

In molecular biology, a *codon* consists of three nucleotides, and each codon determines one amino acid. A sequence of amino acids forms a protein, which produces the functional phenotype of an organism. A single nucleotide substitution on a codon makes it change to another one. Due to the redundancy of genetic codes by degeneration, different codons may encode the same amino acid (e.g., codons *AAA* and *AAG* both code for amino acid lysine). Thus, a nucleotide substitution on a codon may be *synonymous*, i.e., no amino acid replacement. While two different codons generated by a nucleotide substitution can produce different amino acids, this nucleotide substitution is a *nonsynonymous* (amino acid) change. To characterize each site on a codon, in particular, for a codon ε, if $f_\varepsilon(i)$ $(i = 1, 2, 3)$ denotes the fraction of nonsynonymous single-nucleotide substitutions among all possible single-nucleotide substitutions at site i, therefore, the number of nonsynonymous sites on codon ε is $\sum_{i=1}^{3} f_\varepsilon(i)$, and subsequently, the number of synonymous sites on codon ε is $3 - \sum_{i=1}^{3} f_\varepsilon(i)$ [8].

Biologists compare two homologous protein-coding nucleotide gene sequences from related species. These two relevant sequences carry similar genes, i.e., are homologous. However, there can be differences at some nucleotide loci as a result of evolution. Some of these differences on the two gene sequences may result in generating different amino acids for encoding proteins, i.e., are nonsynonymous substitutions, and some of them may not modify the proteins, i.e., are synonymous. The differences between two homologous gene sequences are counted by pairwise comparison of codons. Specifically, the number of nonsynonymous nucleotide substitutions is denoted by M_a, and that of synonymous nucleotide substitutions is M_s. Further, the total number of nonsynonymous (synonymous, resp.) sites for an entire gene sequence is calculated by summing up all the numbers of nonsynonymous (synonymous, resp.) sites on each codon. For the two comparative gene sequences, N_a means the average number of nonsynonymous sites of two sequences. Similarly, N_s is obtained as the number of synonymous sites. Therefore, the nonsynonymous substitution rate $k_\mathrm{a} = M_\mathrm{a}/N_\mathrm{a}$ is the number of observed nonsynonymous substitutions divided by the total number of such type of changes that these sequences *are capable of.* This is a metric of how much evolution has occurred in protein sequences normalized by all possible genetic variations between the two species. Rate $k_\mathrm{s} = M_\mathrm{s}/N_\mathrm{s}$ is the number of observed synonymous changes divided by the total number of such changes that the sequences are capable of. This metric measures the “background” rate of “silent” genetic evolution without phenotypical improvement between the two species.

Therefore, the ratio $k_\mathrm{a}/k_\mathrm{s}$ quantifies the rate of evolution by stating efficient evolutionary changes in relation to silent background evolutionary changes. This ratio also reflects the selection pressure on the evolution of organisms. In the case of $k_\mathrm{a}/k_\mathrm{s} > 1$, fixation of nonsynonymous substitutions is faster than that of synonymous substitutions, which means that *positive selection* fixes amino acid changes faster than silent ones. While mostly one finds $k_\mathrm{a}/k_\mathrm{s} < 1$, the case where deleterious substitutions are eliminated by *purifying selection* (negative selection), and the rate of fixation of amino acid changes is reduced. If $k_\mathrm{a} = k_\mathrm{s}$, the fixation of these two types of changes are at the same rate. Measuring a large $k_\mathrm{a}/k_\mathrm{s}$ ratio suggests that adaptive genetic variations have been generated and fixed at a high rate.

3 Measuring Rate of Evolution in EC

Inspired by the $k_\mathrm{a}/k_\mathrm{s}$ measurement on the rate of sequence evolution in biology, we define the rate of evolution and propose a measurement of it for EC systems. An EC system better capable of evolution can generate efficient adaptation under selection pressure, so it has a *potential* to improve fitness. Apparently, this capability or potential is less observable than fitness itself. Since fast evolution is caused by generating adaptive variations at a high rate we can focus on the adaptive genetic changes underneath the phenotypical fitness to investigate the evolutionary progress. Here, we define the *rate of evolution* R_e as the rate of adaptive genetic changes being accepted into an EC system. Since selection acts at the phenotypical level, the adaptation of a genetic change to its environment can be determined by its acceptance into the evolutionary population. Some changes that are able to improve the adaptation will be accepted, i.e., nonsynonymous substitutions, while other attempted deleterious changes will be eliminated. Some silent changes will be accepted as synonymous substitutions without experiencing selection pressure on phenotypical improvement. Dividing the rate of adaptive substitutions by the rate of synonymous substitutions can quantify the rate of adaptive evolution in an EC system. Therefore, if selection favors the innovated adaptive genetic changes at a high rate relative to the background rate, we say that this EC system has a high rate of evolution.

As a case study, we first utilize a tree-based GP system to implement this idea because GP individuals possess similar features to gene sequences. For example, for a GP tree in our case, genetic changes can be nonsynonymous as in biological systems, which lead to representing different functions, or synonymous, which keep the encoded functions unchanged. We calculate the numbers of substitutions and divide them by the "sites" for a GP system to obtain the two types of rates. Here, we measure the rate of evolution for a GP system of each generation. Specifically, before establishing a generation t, standard mutation and crossover, limited to subtree replacement, are applied to the individual trees in a GP population of generation $t-1$. Truncation tournament selection is then performed on both the parents and offspring to form the next generation t. In such an iteration, we define the rate of evolution $R_\mathrm{e}(t)$ of generation t by observing the genetic changes and their acceptance into the population.

It is well known that changes to a GP tree may be silent due to the existence of neutral *intron* codes [2]. That is, syntactic changes to a tree may not lead to functional changes. Therefore, after mutation or crossover of the trees, these subtree replacements are either nonsynonymous or synonymous. For each individual tree i, if a change is silent, the value of nonsynonymous change $m^i_\mathrm{a}(t)$ is set to 0 and the value of synonymous change $m^i_\mathrm{s}(t)$ is set to 1. In contrast, if a change leads to functional differences, $m^i_\mathrm{a}(t)$ is 1 and $m^i_\mathrm{s}(t)$ is 0. If tree i is not modified from generation $t-1$ to generation t, both $m^i_\mathrm{a}(t)$ and $m^i_\mathrm{s}(t)$ remain 0. After the truncation tournament selection chooses new individuals from both the parents and offspring, a new generation t is established. As a result, the total number of nonsynonymous substitutions $M_\mathrm{a}(t)$ and synonymous substitutions $M_\mathrm{s}(t)$ for the entire population of generation t can be calculated as

$$M_\mathrm{a}(t) = \sum_{i=1}^{S} m^i_\mathrm{a}(t) \ \ , \quad M_\mathrm{s}(t) = \sum_{i=1}^{S} m^i_\mathrm{s}(t) \ \ , \tag{1}$$

where S is the population size. Note that, $M_\mathrm{a}(t)$ and $M_\mathrm{s}(t)$ only count those genetic changes accepted into the population, i.e., substitutions, which have survived through the selection.

As we discussed in the biological $k_\mathrm{a}/k_\mathrm{s}$ ratio definition (Sect. 2), the numbers of nonsynonymous sites and synonymous sites represent the *potential* of the sequence to produce nonsynonymous or synonymous changes, and are used to "normalize" the numbers of substitutions. Here, we adopt a *sensitivity* notion to describe the potential of a GP tree to change its semantic meanings in event of a subtree replacement. Trees have varying sensitivities against subtree replacements, a observation made by Langdon and Banzhaf [6] in research on repeated patterns in tree-based GP systems. We keep a record of all changes to a tree from the beginning of evolution including all attempted subtree replacements, such that the accumulated fraction of these changes being nonsynonymous or synonymous can be regarded as the nonsynonymous sensitivity and synonymous sensitivity of this tree. Specifically, for an individual tree i after initialization, we use $c^i_{\mathrm{a}}(t)$ and $c^i_{\mathrm{s}}(t)$ to denote the accumulated numbers of nonsynonymous and synonymous changes of generation t, respectively, obtained by summing up all the previously recorded changes that have happened to this tree,

$$c^i_{\mathrm{a}}(t) = c^i_{\mathrm{a}}(t-1) + m^i_{\mathrm{a}}(t) \ , \quad c^i_{\mathrm{s}}(t) = c^i_{\mathrm{s}}(t-1) + m^i_{\mathrm{s}}(t) \ , \tag{2}$$

with

$$c^i_{\mathrm{a}}(0) = c^i_{\mathrm{s}}(0) = 0 \ . \tag{3}$$

Therefore, the nonsynonymous and synonymous sensitivities of tree i of generation t can be obtained as follows from the fraction of each type of changes, and these metrics indicate the degree of tree i being changed nonsynonymously or synonymously,

$$n^i_{\mathrm{a}}(t) = \frac{c^i_{\mathrm{a}}(t)}{c^i_{\mathrm{a}}(t) + c^i_{\mathrm{s}}(t)} \ , \quad n^i_{\mathrm{s}}(t) = \frac{c^i_{\mathrm{s}}(t)}{c^i_{\mathrm{a}}(t) + c^i_{\mathrm{s}}(t)} \ . \tag{4}$$

We add up the sensitivities of all individuals in the population to obtain the total nonsynonymous and synonymous sensitivities as the “sites” of the current generation,

$$N_{\mathrm{a}}(t) = \sum_{i=1}^{S} n^i_{\mathrm{a}}(t) \ , \quad N_{\mathrm{s}}(t) = \sum_{i=1}^{S} n^i_{\mathrm{s}}(t) \ . \tag{5}$$

Last, we define the nonsynonymous and the synonymous substitution rates k_{a} and k_{s} of generation t as

$$k_{\mathrm{a}}(t) = \frac{M_{\mathrm{a}}(t)}{N_{\mathrm{a}}(t)} \ , \quad k_{\mathrm{s}}(t) = \frac{M_{\mathrm{s}}(t)}{N_{\mathrm{s}}(t)} \ . \tag{6}$$

The rate $k_{\mathrm{a}}(t)$ measures the rate of generating nonsynonymous adaptive changes. The rate $k_{\mathrm{s}}(t)$ describes the rate of producing neutral changes in an evolutionary process. Without changes at the functional level, these neutral changes will not experience pressure in evolution. Thus, $k_{\mathrm{s}}(t)$ practically provides “clock ticks” for the acceptance of genetic changes in the GP system. Since $k_{\mathrm{a}}(t)$ measures the rate of accepted effective changes, the ratio $k_{\mathrm{a}}(t)/k_{\mathrm{s}}(t)$ represents the “evolutionary distance” in relation to the “evolutionary time”, therefore, the rate of effective adaptation of generation t. Thus, we propose the rate of evolution R_{e} in the GP population of generation t to be

$$R_{\mathrm{e}}(t) = \frac{k_{\mathrm{a}}(t)}{k_{\mathrm{s}}(t)} \ . \tag{7}$$

4 Experimental Results

We calculate R_{e} measurement using GP to solve a benchmark quintic polynomial symbolic regression problem $x^5 - 2x^3 + x$ defined by Koza [5]. Each individual in this GP population is a syntax tree initialized by the method *ramped half-and-half* with maximum depth 6. Candidate functions are evolved toward a target function $f(x) = x^5 - 2x^3 + x$ within interval $[-1, 1]$ by matching a set of sample points. The sample set has 50 real numbers uniformly distributed in $[-1, 1]$. The absolute difference between output and the target $f(x)$ value is the error, and the fitness function is defined as the average error over these 50 samples. The terminal set includes variable x and random ephemeral constants generated from 2001 numbers with equal possibility from $[-1, 1]$ with granularity of 0.001. The four arithmetic operators: $+$, $-$, $\times$, protective $\div$ are used as the function set. We apply random mutation and crossover with probabilities 0.1 and 0.9, respectively, and the maximum mutation subtree depth is 4. Parent individuals and offspring after genetic changes compete through truncation selection with tournament size 4. This GP system has a population size of 4000 evolved for a maximum of 50 generations. A set of 20 cases are used as inputs to a GP tree before and after mutation or crossover, to test each subtree replacement a nonsynonymous or a synonymous one. If all 20 cases produce the same output, subtree replacement applied to this tree is regarded synonymous; otherwise, this tree is considered to have undergone a nonsynonymous change.

A preliminary experiment of this rate of evolution measurement on a single GP evolutionary process can be found in Hu and Banzhaf [4]. Here, we compare R_{e} in different configuration scenarios by varying such parameters as selection size, population size, mutation rate, and crossover rate, to study their effects on evolution acceleration and to verify the effectiveness of our measurement. In each set of experiments, we only change the investigated parameter and hold the others constant. The average fitness, k_{a}, k_{s} and R_{e} are plotted with the average values of 50 successful runs. The method *exponentially weighted moving average* is used here to smooth the curves (smoothing factor 0.1).

4.1 Tournament Selection Size

We increase tournament selection size from 4 to 6 and to 8 (Fig. 1). It is generally accepted that a larger tournament selection size generates greater survival pressure, and thus can maintain better fitness in a population. It can be seen that the population under tournament selection size 8 has the best average fitness. However, due to a higher selection pressure, fewer innovative individuals are accepted, so the population under tournament size 8 has the lowest nonsynonymous substitution rate k_{a}. In contrast, relatively more silent changes are accepted with a larger tournament selection size. This also concurs a recent prediction by Luke and Panait [7] that neutral codes bloat in GP is caused by the pressure of improving fitness. Therefore, the rate of evolution R_{e} decreases as the tournament size increases. These results show that higher selection pressure slows down the rate of accepting genetic variations.

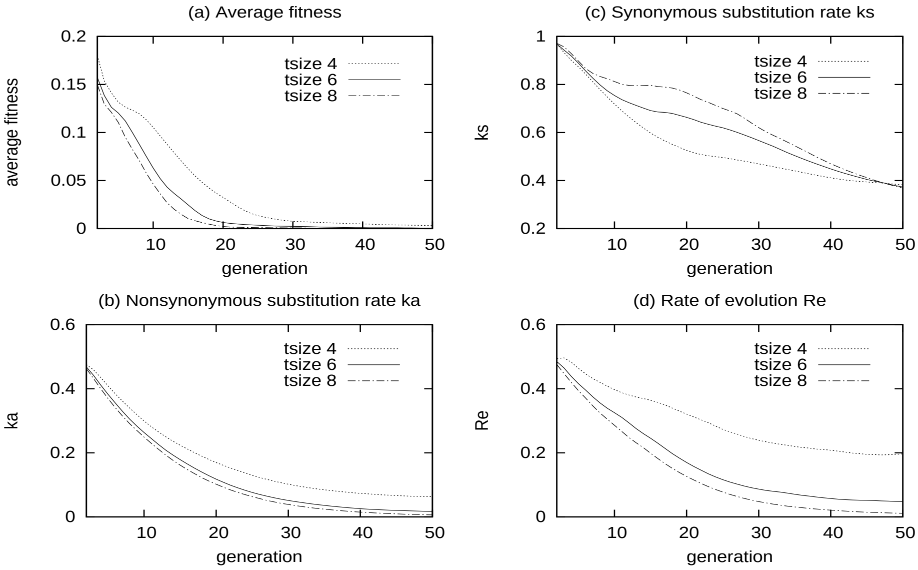

Fig. 1. Rate of evolution with different selection pressure

4.2 Population Size

We test the GP system with the population sizes 200, 2,000 and 20,000 (Fig. 2). Observe that a larger population is better at searching and maintaining the average fitness. All three nonsynonymous substitution rates k_a with different population sizes are quite close, which indicates that, although larger populations offer a larger amount of adaptive individuals to be generated and accepted, their rates in this static symbolic regression problem are nearly the same as smaller populations. Further, a larger population accepts synonymous genetic changes at a slower rate, which is an expected result of a slower propagating speed of dominant individuals. It can be observed that a larger population has a slightly higher R_e at the early stage of the search process but slows down when the target individual becomes dominant in the population. These differences are quite small, however, for this static optimization problem. So we believe that, although a larger population offers more chances of innovating adaptation, under the same environment and selection pressure, a larger population does not have a real advantage in improving the rate of evolution. It can be seen further that, the population with size 200 has the most drastically changing rates, accepting genetic changes at a fairly high rate even around generation 50 (see also the average fitness chart).

4.3 Mutation Rate

The mutation rate is set to 0.3, 0.6, and 0.9 when the crossover rate is fixed to 0.1 (Fig. 3). In our simulations, we only collect successful runs which can

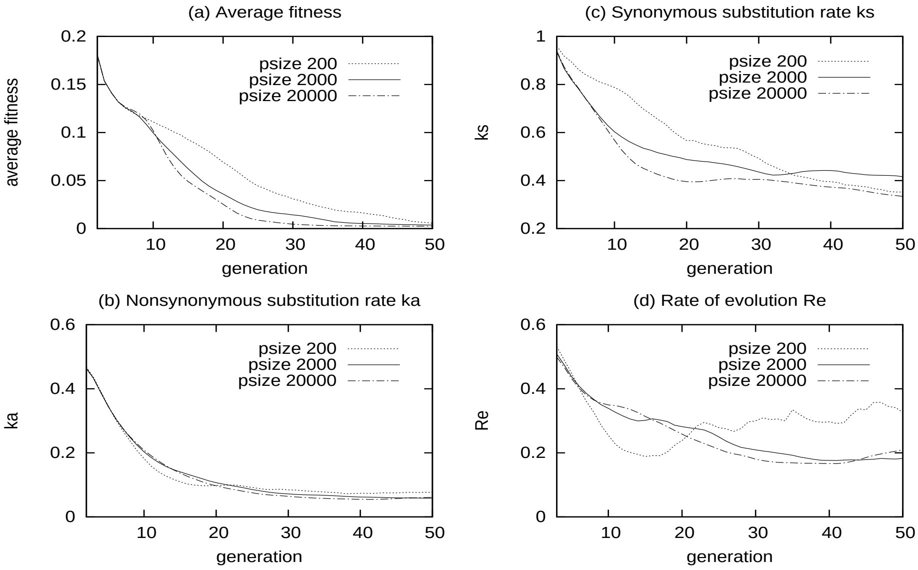

Fig. 2. Rate of evolution with different population sizes

reach the target function within 50 generations. A Population with a higher mutation rate is more likely to succeed. As we observed that, the percentages of successful runs with mutation rates 0.3, 0.6, and 0.9 are 16%, 22%, and 30%. However, despite different success likelihoods, various mutation rates do not show significant differences in the rate of improving the average fitness solving this problem. In our rate of evolution measurement, it can be observed that, a higher mutation rate results in a higher nonsynonymous substitution rate k_{a} and a lower synonymous rate k_{s}, and thus, a higher evolution rate R_{e}. These results show that a higher mutation rate can accelerate evolution but also brings in more noise at the end of evolution (Fig. 3 (d)). Moreover, this simulation supports a general agreement that mutation can maintain good population diversity.

4.4 Crossover Rate

In this set of simulation, we fix the mutation rate at 0.1 and increase the crossover rate from 0.3 to 0.6, and to 0.9. In Fig. 4, similar to varying mutation rates, we can see that investigating fitness development is not sufficient for the effectiveness of crossover rate on the rate of adaptive evolution. In our measurement, it is observed that a larger crossover rate provides more adaptive genetic changes, i.e., a greater k_{a}, and subsequently a higher rate of evolution R_{e}. However, the differences between mutation and crossover operations are their effects on synonymous substitution rate k_{s}. That is, increasing crossover rate can result in a higher synonymous rate, which implies that crossover contributes more to the neutral evolution than mutation.

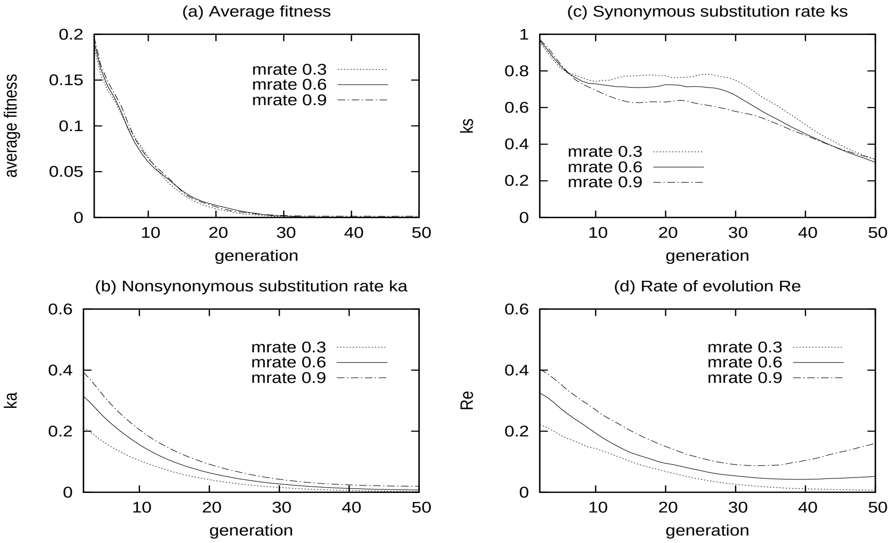

Fig. 3. Rate of evolution with different mutation rates

5 Conclusion and Future Work

In this paper, we introduced the equivalent of a biological measurement of the nonsynonymous to synonymous substitution ratio k_a/k_s. The experimental applications show the ability of this measurement to capture the rate of generating efficient genetic variations in an EC system. Therefore, we believe that the rate of evolution should be better defined by looking beyond fitness and should be measured by the rate of adaptation being generated and accepted. Moreover, some observations are drawn in the simulations that, in the truncation selection scheme, tournament size, mutation rate, and crossover rate are directly related to the rate of evolution, while population size has an indirect relation.

The R_e measurement can be extended in different ways. First, this measurement can be considered to help design adaptive population size in EC models. Through visualizing the rate of evolution at different stages, adaptive population size can be decided to provide effective diversity. Therefore, population size can be chosen systematically rather than empirically. Second, applications of this measurement to various methods in evolutionary computation need to be thoroughly investigated. Third, we propose to use this method for research on quantification of evolvability since it can reflect the evolutionary capabilities of an artificial evolutionary system.

6 Acknowledgements

The authors acknowledge support from NSERC Discovery Grant under RGPIN 283304-07.

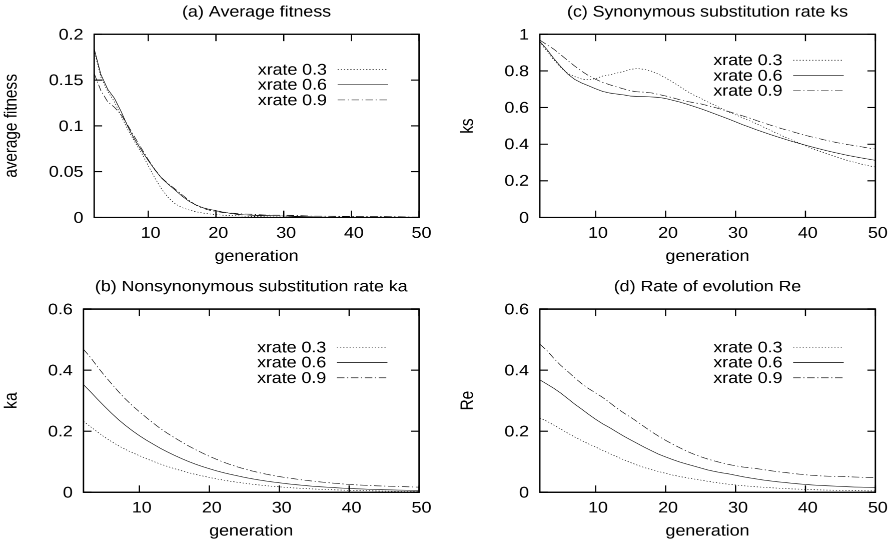

Fig. 4. Rate of evolution with different crossover rates

References

1. W. Banzhaf, G. Beslon, S. Christensen, J. A. Foster, F. Kepes, V. Lefort, J. F. Miller, M. Radman, and J. J. Ramsden. From artificial evolution to computational evolution: A research agenda. *Nature Reviews Genetics*, 7(9):729–735, 2006.
2. W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone. *Genetic Programming: An Introduction On the Automatic Evolution of Computer Programs and Its Applications*. Morgan Kaufmann Publishers, San Francisco, CA, 1998.
3. M. A. Bedau and N. H. Packard. Measurement of evolutionary activity, teleology, and life. In *Artificial Life II*, pages 431–461. Addison-Wesley, Redwood City, CA, 1992.
4. T. Hu and W. Banzhaf. Measuring rate of evolution in genetic programming using amino acid to synonymous substitution ratio k_a/k_s. To appear in *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, Atlanta, GA, 2008.
5. J. R. Koza. *Genetic programming II: automatic discovery of reusable programs*. MIT Press, Cambridge, MA, 1994.
6. W. B. Langdon and W. Banzhaf. Repeated patterns in tree genetic programming. In *Proceedings of the 8th European Conference on Genetic Programming*, pages 190–202, Lausanne, Switzerland, 2005.
7. S. Luke and L. Panait. A Comparison of Bloat Control Methods for Genetic Programming. *Evolutionary Computation*, 14(3):309–334, 2006.
8. T. Miyata and T. Yasunaga. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16(1):23–36, 1980.
9. Z. Yang and J. P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15(12):496–503, 2000.