特別研究報告

題目

遺伝的アルゴリズムを採用した自動プログラム修正における個体生成時間の定量的評価

指導教員

楠本 真二 教授

報告者

皆森 祐希

令和 5 年 2 月 7 日

大阪大学 基礎工学部 情報科学科

令和 4 年度 特別研究報告

遺伝的アルゴリズムを採用した自動プログラム修正における個体生成時間の定量的評価

皆森 祐希

内容梗概

ソフトウェア開発において，デバッグは開発工数の半分を占める作業といわれている．自動プログラム修正は，バグを含むプログラムに変更を加えることで用意したテストを通過するプログラムを出力する手法であり，デバッグの工数を削減することが期待されることから研究が盛んにおこなわれている．自動プログラム修正の手法として，遺伝的アルゴリズムに基づいて修正を行うものがある．ここで，自動プログラム修正における遺伝的アルゴリズムは，目的のプログラムが得られるまでプログラム文の挿入，削除，置換及び交叉を行う手法である．現状の自動プログラム修正における課題の 1 つとして，1 つのプロジェクトに対する修正に長い時間がかかる点があげられる．しかし，現状の自動プログラム修正ツールにはプロジェクトの修正時間のうち個体生成に対する計測機能が存在しない．そこで，本研究では遺伝的アルゴリズムを採用した自動プログラム修正ツールのひとつである kGenProg に各個体の生成にかかった時間を計測する処理を追加した．さらに得られた個体情報をビルドの成否および解であるかどうかに分類しそれらを定量化した．また全体の生成時間のうち解となるプログラムの生成経路にかかった時間およびビルドに失敗した個体に費やした時間を計算することで，自動プログラム修正の時間的効率性を調査した．

主な用語

自動プログラム修正， 時間計測，可視化， JSON

目次

図目次

表目次

1　はじめに

ソフトウェア開発におけるデバッグ作業は費用および時間的な点で多くのコストを必要とする．ある研究によると，ソフトウェア開発にかかるコストのうち，50% 以上をデバッグが占めるという結果が出ている [?, ?]．自動プログラム修正 (APR) は，人の手を介さずにソースコード中に含まれるバグを自動的に取り除く技術であり，盛んに研究が進められている [?, ?]．APR の実用化に向けて，ここ 10 年で数多くの研究がなされており，そのうち時間的なコストを削減する取り組みが数多く行われている [?]．TODO：時間的コスト削減についての追加の論文調査しかし，先行研究においては個体の生成時間に関する研究は多くない．そこで，本稿では，既存の APR ツールを拡張して個体の生成時間を計測し，筆者が考案した指標を実際のバグに対して計算することで時間的コストを調査する．

以降，2 章では APR の課題本研究を行う契機となった研究の説明および研究のための予備知識について説明する．3 章では本研究における提案手法について論じる．4 章では今回提案した評価指標とその具体例について論述する．5 章では実験の概要と結果を提示する．6 章では実験から得られた結果による考察について述べる．7 章では本研究の妥当性への脅威について論じる．8 章では今後の課題について説明し，最後に 9 章で総括を述べる．

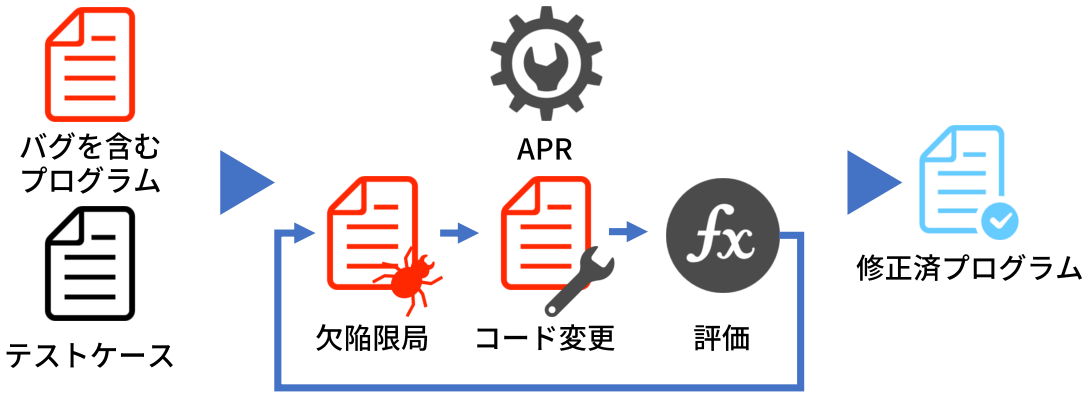

図 1　自動プログラム修正：APR

2　準備

2.1　自動プログラム修正 (APR)

APR は，テストケースとバグを含むプログラムを入力として，計算機が自動的にプログラムのバグ修正を行う技術である．この時，出力は入力したプログラムを修正したプログラムである．APR は大きく探索ベース[*1]と意味論ベース[*2]の 2 つの手法に分類することができるが，本研究では，探索ベース APR 分野のブレイクスルーとなった GenProg [?] の採用する生成と検証 [?] に基づく手法に重点を置く．

この手法では，図 1 のように，対象となるプログラムにおけるバグの位置を特定する欠陥限局を行い，限局した箇所に変更を加えた後ビルドとテストを実行して修正尺度の評価を行う．

2.2　遺伝的アルゴリズム

遺伝的アルゴリズム (GA) は各世代で個体を選択し，それらに変異，交叉などの操作を加えることでより強い個体を生成する生物の進化に基づくアルゴリズムである．これらの個体から以下の遺伝子操作を行うことで，次の世代の個体の集合を生成する．

[*1] Heuristics-based
[*2] Semantics-based

選択 …個体のうちから何らかの関数による適応度 (APR ではテストスイートの通過率) に応じて選択

変異 …個体の遺伝子を変異させる (APR ではコードの一部を変更)

交叉 …複数の遺伝子の一部分を交配させて新しい遺伝子を生成

これらの遺伝子操作のうち，変異において APR では以下の操作を対象のコードに対して行う．

挿入 …選択したコード近辺への別コードの追加

置換 …選択したコードの別コードへの書き換え

交叉 …選択したコードの削除

遺伝的アルゴリズムの具体的な説明として，図 2 のような遺伝について考える．なお，図中の数字はテストスイートの通過率 (以下，単に通過率と表す) を表し，もっとも左の個体群を第 1 世代として，右に行くほど世代が進んでいるものとする．
第 1 世代においては，通過率 0.5 の個体を選択したもの，通過率 0.3 の個体に変異を起こしたもの，そして通過率 0.4 と 0.3 の個体を交叉したものの 3 つの個体を第 2 世代の個体として生成している．同様に，第 3 世代の個体は，第 2 世代の個体のうち，通過率 0.7 の個体の選択およびこの個体と通過率 0.5, 通過率 0.6 の個体との交叉による個体となっている．ここで，第 3 世代の個体に通過率 1.0 の個体が存在するのでここでアルゴリズムを終了する．

2.3 既存の APR ツールの課題

既存の APR ツールは多くの課題が解決されておらず [**?**]，まだまだ実用には程遠い段階である．具体的には，修正後のプログラムの可読性が低い [**?**]，どがあげられる．
その中で今回主に取り上げる課題として，1 回の修正実行にビルドやテスト実行といった多くの時間的コストがかかる点 [**?**] があげられる．また，先行研究においては，ビルドとテストにかかる時間を削減する研究が行われているものの [**?**] ，個体の生成そのものにかかった時間を計測する機能に関する研究はまだ発展途上である．

2.4 時間的コスト調査の先行研究

Ghanbari [**?**] らはソースコードレベルの APR に加えて，さらにバイトコードに APR を施す PraPR を提案し, 既存の APR の性能を飛躍的に向上させた．この研究において, 提案した PraPR を Defects4J の Chart および Closure バグに対して時間的コストを計算した．結果として，Closure バグでは有効なパッチの数が Chart バグに比べて 10 倍生成されたものの，時間的コストは 20 倍かかった．また，古藤

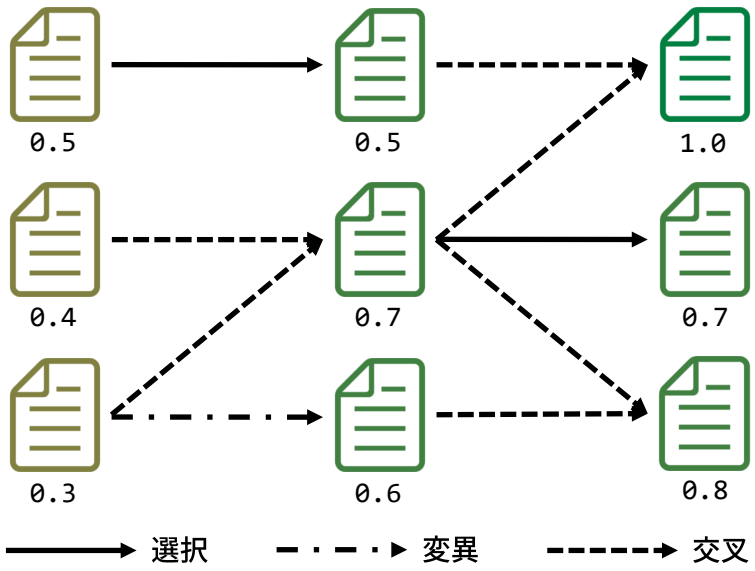

図 2　遺伝的アルゴリズムの例

[?] らは欠陥限局を用いて変更コード片を動的に切り替える手法を提案しビルドに費やす時間を従来手法に比べて 89% から 46% に削減することができ，結果として APR 全体の修正時間の削減につながった．これらの研究ではプログラム修正全体やビルド時間に対する時間的コストの調査自動プログラム修正の時間的コストについてより詳らかに調べてみようと思い，本研究をするに至った．

3　提案手法

前の章でも述べた通り，APR ツールにおける個体の生成にかかった時間に重点を置いた研究はまだ少ない．そこで，本研究ではの遺伝的アルゴリズムを採用した GenProg [?] 系の APR ツールの個体処理部分に時間計測用のコードをはさんで，記録された生成時間を独自の指標で評価する手法を提案する．

4　評価指標とその実装

4.1　評価指標

4.1.1　STR

APR を実行するにあたり，修正に成功したプロジェクトに対して，その解となる個体に関係する経路の全体の生成時間に占める割合を求める指標として，**STR**(Solution Time Ratio, 解時間比率) を次の式で定義する．

$$\mathrm{STR} = \frac{\text{解となる個体の経路の総生成時間}}{\text{すべての個体の総生成時間}} \tag{1}$$

ここで，解となる個体の経路の集合は

1. まず解である個体を選択し，集合に入れる
2. 集合内の全個体に対してその親を求め，それらを集合に入れる
3. 集合に変化がなくなるまで 2 を繰り返す

のように求められる．STR を計算する目的として，プログラム修正において成功に必要な個体の生成が時間的にどの程度の割合を占めているのかを定量的に求めることがあげられる．そのため，STR の値は大きい方が好ましい．

4.1.2　FTR

一方で，すべてのプロジェクトに対してビルドに失敗する個体がすべての個体の生成時間に占める割合を求める指標として，**FTR**(Failure Time Ratio, 失敗時間比率) を次の式で定義する．

$$\mathrm{FTR} = \frac{\text{ビルドに失敗した個体にかかった総生成時間}}{\text{すべての個体の総生成時間}} \tag{2}$$

FTR を計算する目的として，プロジェクト内のプログラムの修正にかかる時間のうちがどの程度の割合を占めるかを定量的に測定し，その傾向を知ることがあげられる．そのため，一般的に FTR の値は小さい方が望ましいとされる．

4.2　STR・FTR の具体的な計算例

先ほど定義した値を求めるための具体的な例として，図 3 の生成木をもつ修正結果について考える．この生成木は，kGenProg の実行結果を記した JSON ファイルをツリー状に表示する Macaw [**?**] を参考に描画した[*3]．ここで，縦方向は世代を表しており，下に進むにつれてより新しい世代を表す．また，

[*3] ビルドに失敗した個体集合を表す X 印における数字の意味合いとして，この例では生成時間の和を表すが Macaw においては個体の数を指すなどの細かい違いがある点に注意

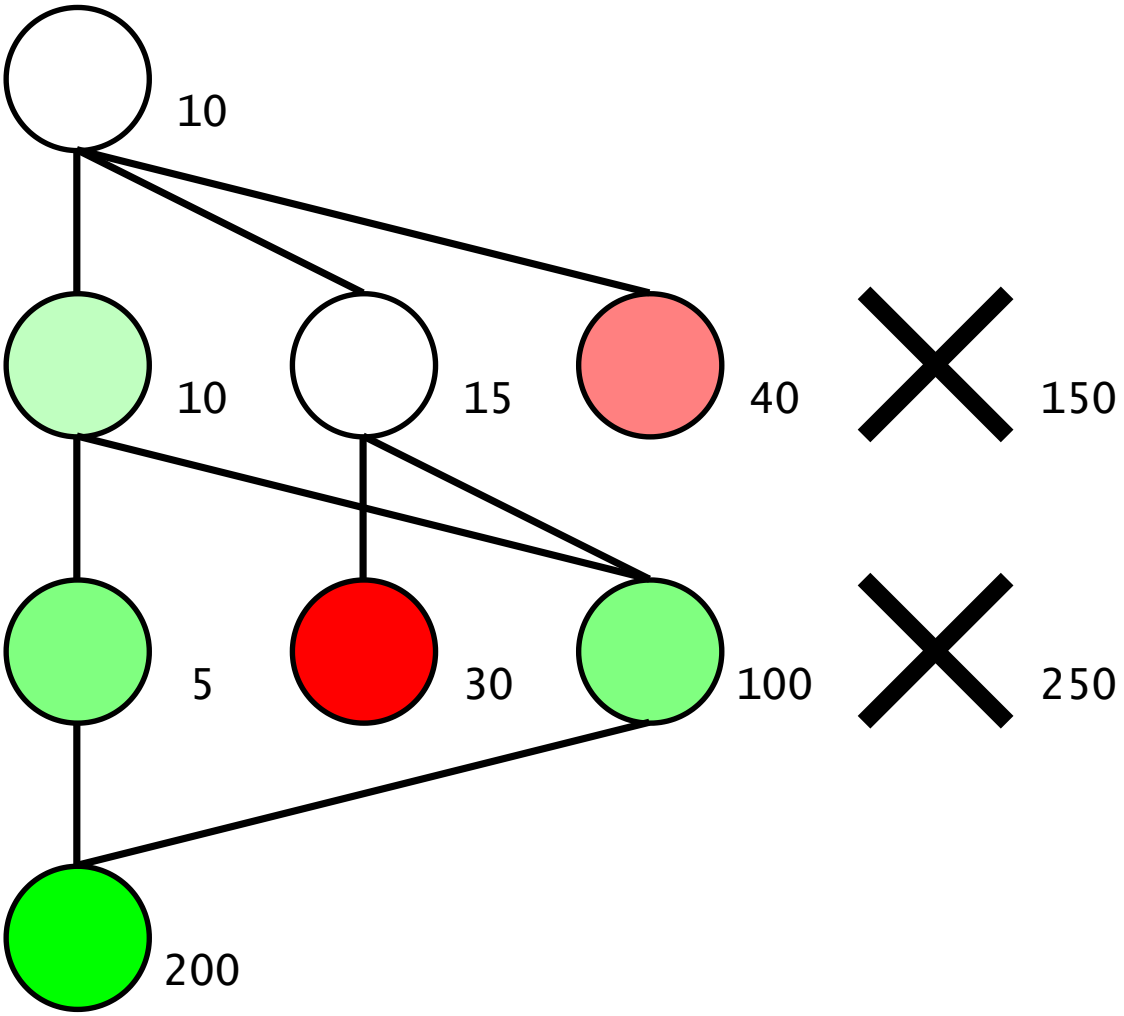

図 3　生成結果の例

横の列は一つの世代におけるすべての個体を表す．円及び X 印はそれぞれビルドに成功した 1 つの個体，その世代でビルドに失敗したすべての個体を表す．円の色は個体の Fitness(全体のテストケースに占める期待通りのテスト結果が得られた割合) を表し，緑に近いほど高い Fitness を右下の数字はその個体あるいは個体の集合の生成時間を表す．なおこの例におけるプログラムの総生成時間は 800 である．この時，STR は図 4 で表される部分の時間 $(10+15+5+100+200) \div 800 = 0.4125$ となり，FTR は X 印で示されたすべての時間を足した値を総生成時間で割ったもの，すなわち $(150+250) \div 800 = 0.5$ と求められる．

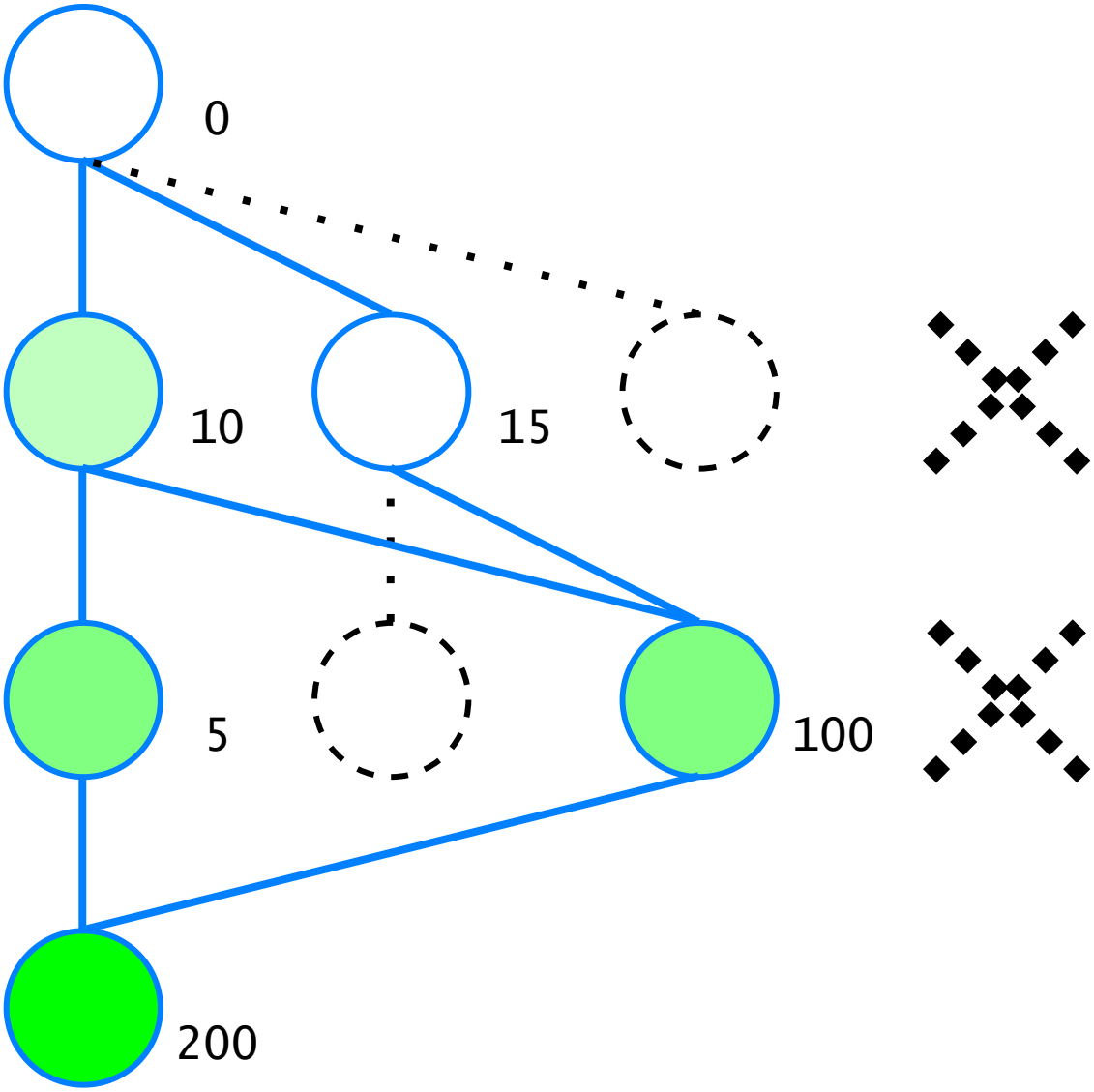

図 4　生成結果の例

4.3　APR ツールへの時間計測機能実装

提案手法では，既存の APR ツールである kGenProg [?] を拡張して実装する．ソースコード内部のふるまいは変えずに，`Mutation` クラスと `RandomCrossover` クラスに時間計測を行うコードを挿入した．具体的には，`org.apache.commons.lang3.time.StopWatch` クラスをインポートし，各個体を生成するループの最初に `StopWatch.createStarted()` メソッドを呼び出すことによって時間計測を開始，処理を終えた後に `getTime()` メソッドを呼び出すことで生成時間を取得し，それを個体情報に格納する．この際，kGenProg に付属している JSON ファイルの出力オプションをオンにすることで，プ

ロジェクトにおける個体の解析を可能にする．

次に，JSON ファイルを Python で記述したプログラムを用いて処理し，各個体の ID(通し番号)・生成時間・Fitness(ただしここでは ID に対応する個体がビルドに失敗した場合-1 を格納する) の情報を取得した後，その情報をもとに STR と FTR を計算する．

具体的には，出力となる JSON ファイルを読み込み，そのプロジェクトで生成された個体の時間を 1 つずつ取得する．この時，Fitness の値で追加の処理を行う．例えば Fitness が-1 であればビルドに失敗した個体であるので失敗時間 (FTR を計算する際の分子) に加算する．また，Fitness が 1 であれば，テストケースを満たす解となる個体であるので 4.1.1 で挙げた手順で解となる個体の親を求める．プログラム中では再帰的なアルゴリズムを用いている．

5　実験

5.1　概要

本章では，GA を採用した APR ツールである kGenProg [?] を用いて，Defects4J [?] の Lang プロジェクトの Lang1〜Lang44 を対象に自動プログラム修正を行った．そのうち，ビルドに成功し，かつ解を得ることができた Lang6, Lang22, Lang25 および Lang39 の 4 つのプロジェクトを対象として先ほど定義した STR と FTR を計算し，その値を確認する．なお，生成時間には不確定性があるため各プロジェクトごとに APR を複数回実行している．表 1 に APR ツール実行時の設定を，表 2 に実行回数や解に至るまでの総個体数など，各プロジェクトに対する実験の条件を示す． ここで，サンプル数がプロジェクトによって異なるのは，1 回のプログラム修正にかかる総時間が異なるためである．

5.2　実験結果

5.2.1　Lang6

図 5 に Lang6 における STR と FTR の箱ひげ図を，表 3 に STR と FTR の平均 (以降 AVG)，最小値 (以降 MIN)，第 1 四分位数（以降 1Q)，中央値 (以降 MED)，第 3 四分位数 (以降 3Q)，最大値 (以降 MAX) の各データを示す.

データからわかることとして，修正を完了するまでに生成した個体の数が 8 つと比較的少ないことも

表 1　APR ツールの設定

項目	値
実験題材	Defects4J Lang6, Lang22, Lang25, Lang39
題材数	4
乱数シード	2
実験環境	Corei5-1240P 16GB mem

表 2　各プロジェクトの詳細

プロジェクト名	Lang6	Lang22	Lang25	Lang39
到達世代数	1	7	1	4
個体数	8	624	38	300
ビルド失敗個体数	7	511	36	274
サンプル数	100	15	70	40

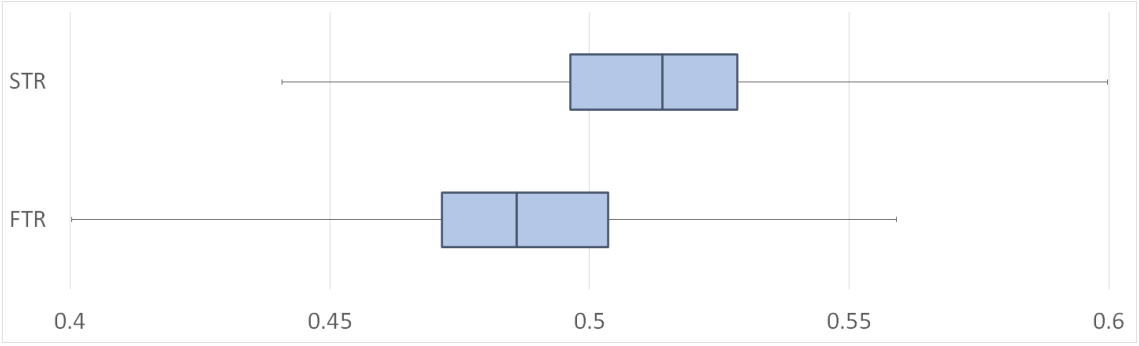

図 5　箱ひげ図：Lang6

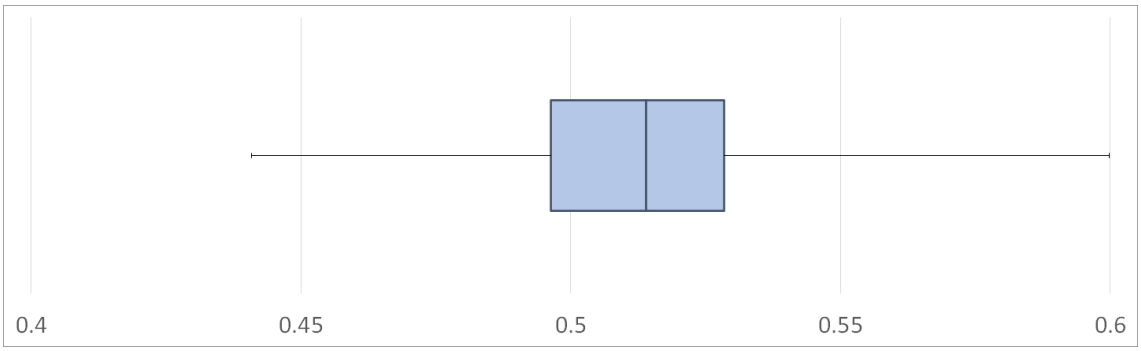

図 6　STR の箱ひげ図：Lang6

あってか，STR，FTR の値はいずれも 0.5 程度となった．この値は，解となる個体に関係する生成時間およびビルドに失敗した個体の生成時間が全体の生成時間のおよそ半分程度であることを意味する．

5.2.2　Lang22

次いで，図 8 に Lang22 における STR と FTR の箱ひげ図を示す．この結果からわかることとして，他のプロジェクトに比べてパッチの生成に多くの時間および操作を要した．そのため，他のプロジェクトに比べると STR，FTR のいずれの値も小さくなっている．

表 3　Lang6 プロジェクトの STR, FTR

評価指標	AVG	MIN	1Q	MED	3Q	MAX
STR	0.5117	0.4409	0.4964	0.5140	0.5284	0.5997
FTR	0.4883	0.4003	0.4716	0.4860	0.5036	0.5591

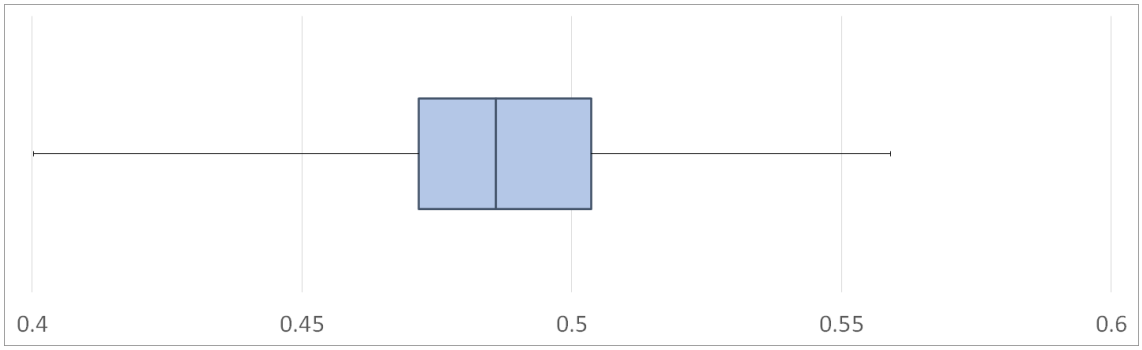

図 7　FTR の箱ひげ図：Lang6

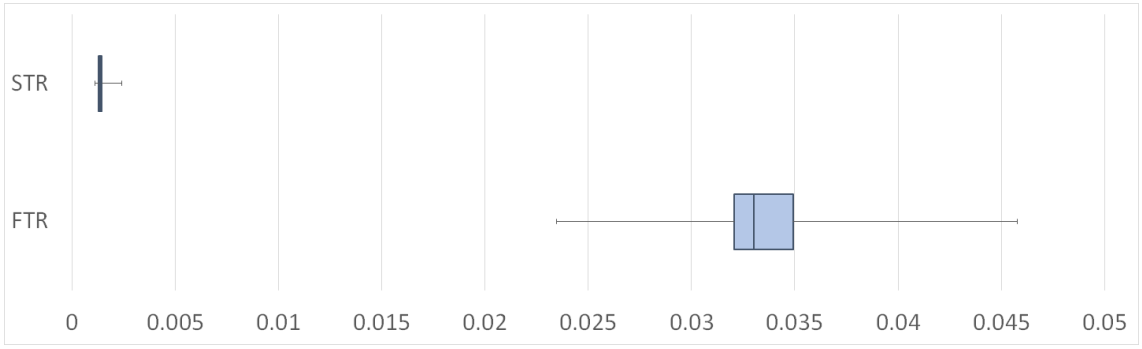

図 8　箱ひげ図：Lang22

5.2.3　Lang25

次に，図 11 に Lang25 における STR と FTR の箱ひげ図を示す．このプロジェクトにおける特徴として，FTR の値が平均して 0.8 程度と高い数値を示している点があげられる．

表 4　Lang22 プロジェクトの STR, FTR

評価指標	AVG	MIN	1Q	MED	3Q	MAX
STR	0.001476	0.001096	0.001291	0.001354	0.001458	0.002418
FTR	0.03432	0.02346	0.03208	0.03304	0.03492	0.04578

表 5　Lang25 プロジェクトの STR, FTR

評価指標	AVG	MIN	1Q	MED	3Q	MAX
STR	0.09386	0.07261	0.08936	0.09106	0.1070	0.1177
FTR	0.7986	0.7371	0.7725	0.7989	0.8053	0.8512

図 9　STR の箱ひげ図：Lang22

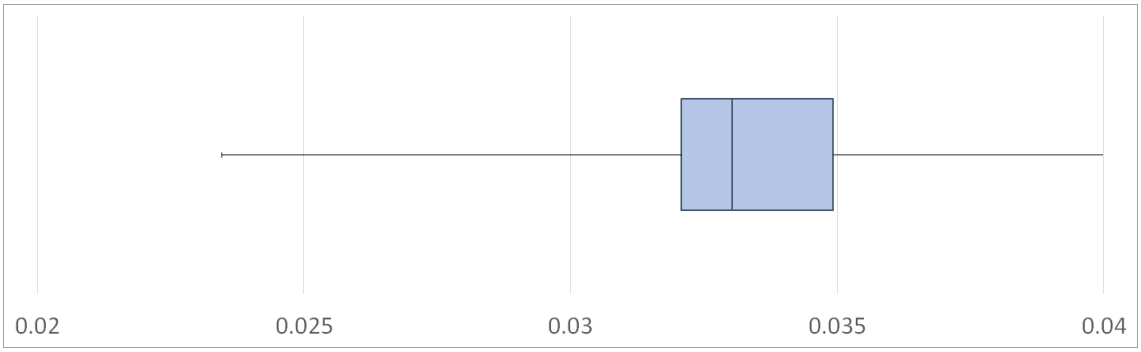

図 10　FTR の箱ひげ図：Lang22

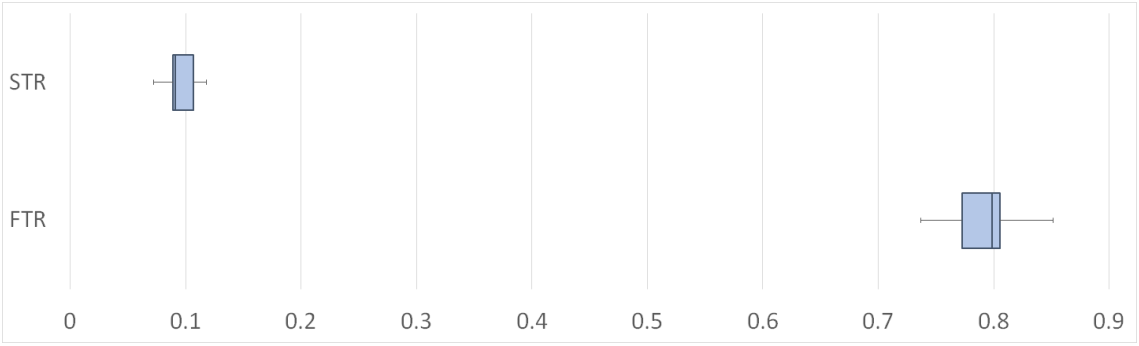

図 11　箱ひげ図：Lang25

5.2.4　Lang39

最後に，図 14 に Lang39 における STR と FTR の箱ひげ図を示す．このプロジェクトでも，Lang25 ほどではないものの，FTR が 0.6 付近と高い値を得た．

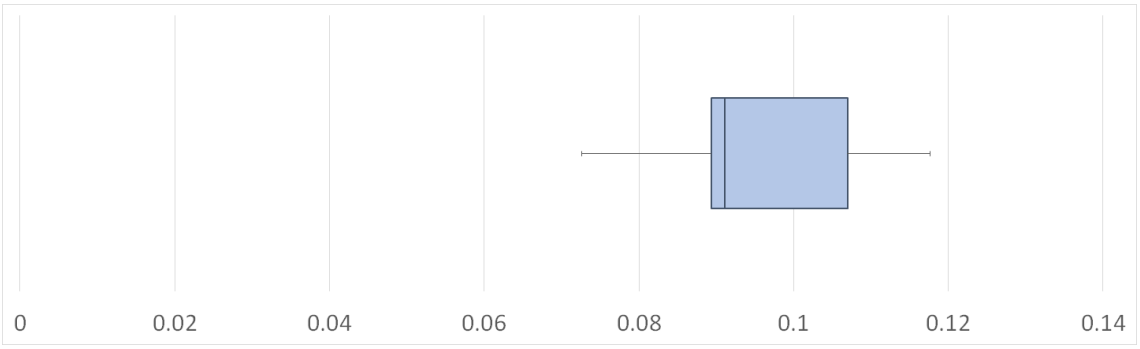

図 12　STR の箱ひげ図：Lang25

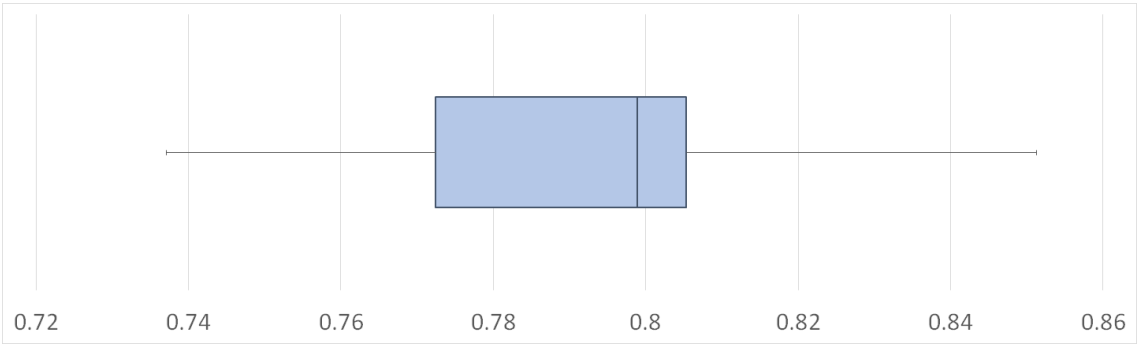

図 13　FTR の箱ひげ図：Lang25

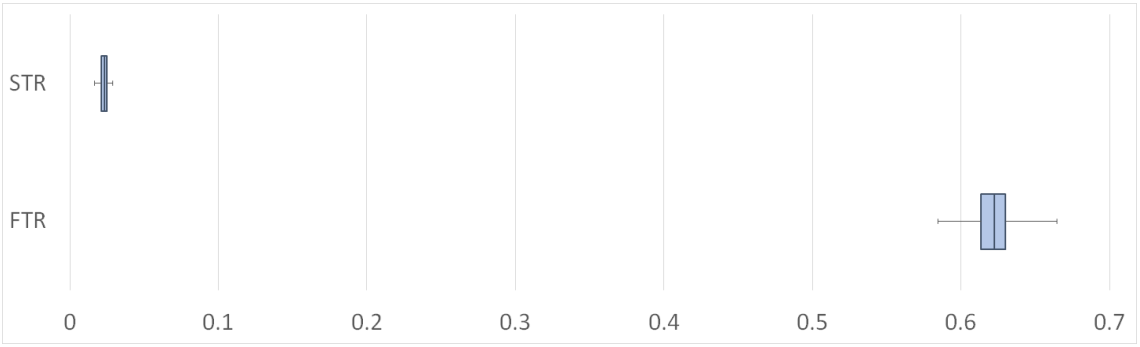

図 14　箱ひげ図：Lang39

表 6　Lang39 プロジェクトの STR, FTR

評価指標	AVG	MIN	1Q	MED	3Q	MAX
STR	0.02406	0.01645	0.02105	0.02326	0.02510	0.02873
FTR	0.6218	0.5847	0.6137	0.6224	0.6299	0.6644

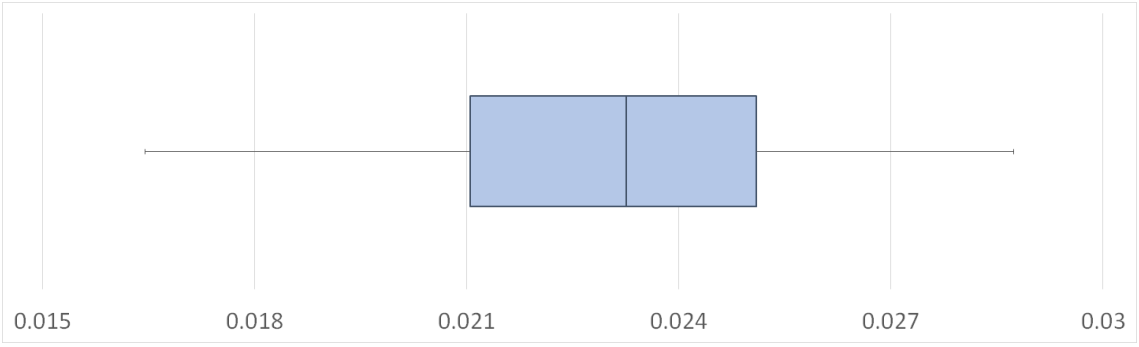

図 15　STR の箱ひげ図：Lang39

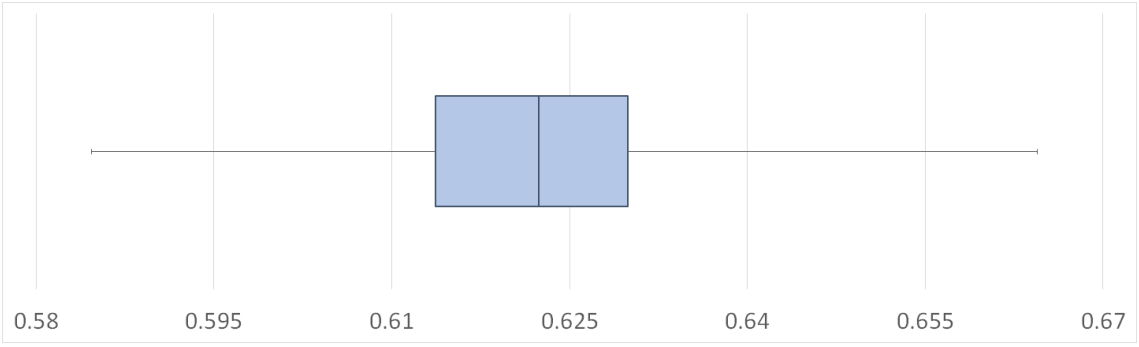

図 16　FTR の箱ひげ図：Lang39

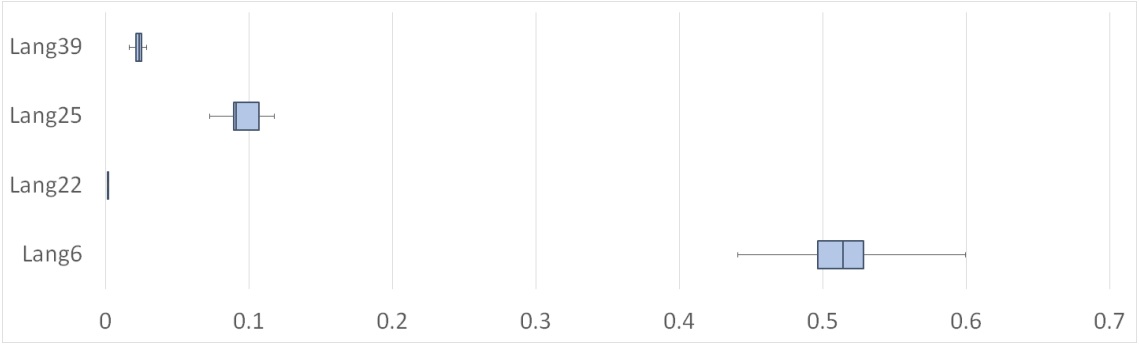

図 17　4 つのプロジェクトにおける STR の箱ひげ図

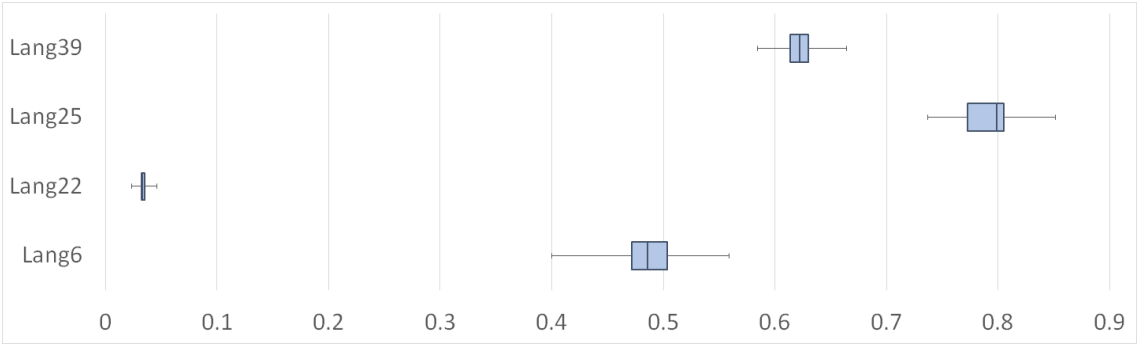

図 18　4 つのプロジェクトにおける FTR の箱ひげ図

6　考察

次に，5 章で行った実験の結果から，各プロジェクトおよび全体的な STR と FTR の傾向について考察する．考察して仕上げる

7　妥当性の脅威

5 章における実験において考えうる妥当性の脅威について論ずる．この実験では内的要因と外的要因に大別される．まず内的要因として，今回実行した環境とは異なる環境において実行した際に算出される STR と FTR の値が異なる可能性がある点があげられる．また，実行時に他のタスクによりメモリが占領されている場合，普段と異なる値が算出されうる点があげられる．

次に外的要因として，大規模なプロジェクトにおいて何万といった個体が生成されたときに STR と FTR の値が大きく異なる今回対象としたプロジェクトは個体数が多いもので 600 程度と比較的少なかった．

8　今後の課題

本研究においては，探索ベースで遺伝的アルゴリズムを採用した kGenProg を対象に個体の生成時間計測を行ったが，

8.1　APR ツールによる生成時間の違いの検証

本研究で用いた APR ツールは探索ベースで遺伝的アルゴリズムを採用した kGenProg を対象に個体の生成時間計測を行ったが，遺伝的アルゴリズムでないほかの探索ベースの APR ツールや意味論ベースの APR ツールにおいて時間計測を行う．

8.2　時間計測を行う対象プロジェクトの拡張

本研究では，Defects4J の Lang バグのうち，4 つのプロジェクトを対象として STR と FTR の計算を行った．今後は，Defects4J のほかのバグやその他の対象について時間研究を行う．

9　おわりに

本稿では，APR ツールに対して個体の生成時間を計測する機能を追加し，実際のバグに対して STR と FTR という 2 つの指標を定義した．さらに実際のバグに対して指標を計算する実験を行うことで時間的コストの評価を行った．この後の文章を完成させる

謝辞

本研究を進めるにあたり，多くの方々からご支援およびご助言を賜りました．

楠本 真二 教授には，本研究を快く快諾し，暖かく見守ってくださりました．心より感謝申し上げます．

肥後 芳樹 准教授には，議論を重ねに重ね，本研究の完成のご支援及び的確なご助言を賜りました．心より感謝申し上げます．

柗本 真佑 助教には，テーマが決まらず途方に暮れていた際，鋭くも的確なご助言を賜りました．深く感謝いたします．

古藤 寛大先輩には，困難に直面した際いつも迅速にご助言を賜り感謝してもしきれません．

本研究に至るまでに，講義，演習等でお世話になりました大阪大学基礎工学部情報科学科の諸先生方に，御礼申し上げます．

最後に，これまでお世話になりました家族，小中高校の教員方，その他すべての方に感謝申し上げます．