International Conference on Computational Science, ICCS 2012

Comparative Analysis of Software Repository Metrics in BioPerl, BioJava and BioRuby

M. Rahmani *, D. Bastola, and L. Najjar

College of IS&T, University of Nebraska at Omaha, 6001 Dodge St., PKI 172, Omaha, NE 68182, USA

Abstract

The open source programming languages, often with a bio- suffix, i.e. BioPerl, BioJava, and BioRuby, have been widely used in bioinformatics and computational biology research. The computational tools written in these languages provide multiple functionalities as the languages make them flexible to create customized analysis and examination of biological data. In this paper, we investigate one of the software quality parameters, maintainability, in BioPerl, BioJava, and BioRuby projects using comment density metric in their source code repositories. Furthermore, we evaluate the “popularity” of these three languages in bioinformatics communities using three other software metrics such as number of committers, commit frequency, and lines of code. To perform this study, source code repositories of these three open source projects have been analyzed from the first release, which covers all the programming activities of the projects from the starting date until July 2011. Our results show BioPerl to be the most popular language among the three languages in open source communities. In addition, investigation on comment density of these three open source projects has shown that BioPerl is the most promising one in terms of future maintainability and quality of the project. The results of this research can be useful for developers in choosing an appropriate language for the development of bioinformatics applications.

Keywords: Bioinformatics Programming Language; BioJava; BioPerl; BioRuby; Software Repository Metrics; Software Maintainability

1. Introduction

The open source languages such as BioPerl, BioJava, and BioRuby have become popular because of special biological data processing and their extensive functionalities to create customized bioinformatics applications. The intent of this research is to present a quantitative comparison of these three bioinformatics languages based on several software repository metrics such as number of committers, commit frequency, Lines of Code (LOC), and comment density. All these metrics are code repository metrics which can be extracted from the version control system of the project. The first three metrics are used as the indicator of “popularity” of a language, which measures the amount of effort spent by developers in the open source community. The later metric, comment density, is used to assess the “maintainability” of the software project.

Based on IEEE Standard Glossary of Software Engineering Terminology, maintainability is the ease with which
a software system or component can be modified to correct faults, improve performance or other attributes, or adapt to a changed environment [1]. Since the amount of comments in a given source code can increase the ease of reviewing code by developers, comment density can be interpreted as an indicator of code quality and maintainability in the project [2]. A review of literature show several research works related to the benchmarking and characterizing of bioinformatics languages [3] [4] where several metrics have been used to compare the bioinformatics languages such as BioPerl, BioJava, and BioPython with its parent Perl, Java, and Python, respectively. The matrices used in those studies include the execution time of a function written by a language, the amount of memory usage for a task, readability of code written by a language, and portability of the final application. Although these works concentrate on different aspects of bioinformatics languages, to the best of our knowledge, there is no work that emphasizes comparing the maintainability and popularity of the open source bioinformatics languages.

Section 2 presents some background knowledge of the languages. Section 3 discusses the research method and the results. Section 4 represents some of the limitations and future work. Finally, Section 5 concludes the paper.

2. Background

BioPerl, BioJava and BioRuby are open source languages that are well-known in bioinformatics research domain but their functionality and usability differ from each other. BioPerl is the oldest, most downloaded package with the largest development community among bioinformatics packages currently in use. BioPerl has a well-designed layout of its document, which allows developers to pursue the functionality of the package more conveniently. Previously, it has been shown that for small programs with 500 lines of code or less, which cover around 90 percent of personal bioinformatics programming requirements, Perl and BioPerl are the dominant languages in use [3]. BioJava is another mature open-source project capable of processing biological data. It contains powerful analysis and statistical routines. It has tools for parsing common file formats, and contains packages for manipulating sequences and 3D protein structures. It enables rapid bioinformatics application development in the Java programming language [5]. BioRuby is another language that contains a comprehensive set of free development tools and libraries for bioinformatics written in the Ruby programming language. It has components for sequence analysis, pathway analysis, and protein modelling and phylogenetic analysis [6]. This language supports many widely used data formats in biomedical informatics and provides easy access to biological databases [7], external programs, and public web services, including BLAST, Entrez, KEGG and GO. BioRuby comes with a tutorial, documentation, and an interactive environment that can be used both in the shell, and in the web browser.

Figure 1 shows the scaled search volume of BioPerl, BioJava, and BioRuby extracted from Google Insights for Search [8] from January 2004 till September 2011. The result is a ratio of a particular search term to the total number of searches done on Google over time. It is seen that BioPerl is a dominant search term in Google. Although the overall search trend for all of these three languages is decreasing with time, the average interest in BioPerl is still highest among the three languages. Additionally, this result represents the overall interest of users to these languages. However, in order to compare popularity and maintainability of these languages, the following four source code repository metrics were analysed.

![Fig. 1. Interest to BioPerl, BioJava, BioRuby over time, presented by Google in September 2011 [8]](image)

**Number of committers:** This metric shows the number of active committers per month, where committer is a developer who makes changes to the software repository. In open source projects, committers determine where the project is headed, both strategically and on a day-to-day basis. They can typically resolve technical problems faster than non-committers, and have high visibility in the user community [9].
Commit Frequency: This metric measures the number of commits by all committers monthly. Every commit made by the software developer records the changes including the addition, deletion and modification to source code made to the software.

Comment density: This is expected to be a good predictor of maintainability and survival of a project [10]. It is represented as percentage of comment lines in a given source code, that is, comment lines divided by total lines of code [11].

Lines of code (LOC): This is a suitable measure of project size and extended functionality of software. LOC, total physical lines, can be calculated for an entire software product. This information was obtained from ohloh.net [12].

3. Research Analysis

The focus of this paper is quantitative analysis of multiple matrices used as the indicator of popularity and maintainability of three open source bio-languages. Each of these two qualitative parameters, i.e. maintainability and popularity, is represented by quantitative metrics extracted from software code repository. While the number of committers, commit frequency, and LOC are used to show project popularity, comment density is used to represent maintainability of the project. Each of these metrics is gathered once a month throughout the life of the project.

As shown (Figure 2A), the BioPerl project has a significantly larger number of LOC. In July 2011, the total LOC for BioPerl, BioJava and BioRuby projects were 717749, 407158, and 109478 lines respectively. One may argue that the LOC is language dependent and it simply shows the amount of programming effort for a project, which may not be a useful evaluation factor to demonstrate total effort spent on a project. We agree that comparison of LOC between these three different languages is not the best approach to compare the amount of effort and consequently its popularity, but we argue that in this case it can be an indicator of highest programming effort for BioPerl project. We refer the reader to research that compares the size of several programs written in different bioinformatics languages which perform the same task [4]. As the major languages used to implement BioPerl, BioJava, and BioRuby are Perl, Java and Ruby respectively, the result of the research in [4] is very interesting and endorses our findings about the popularity of these three bioinformatics open source languages. It shows that Perl required the least lines of code to perform a task compared to other languages. As presented in Figure 2A, BioPerl has the largest LOC. This result together with the information obtained from [4], suggest BioPerl to be a more attractive open source project, and LOC is reflective of more functionality than two other languages.

Figure 2B shows the average number of committers who were active throughout the life of the BioPerl project is 9.73 committers per month. However, this number is 4.16 and 2.08 for BioJava and BioRuby respectively. This shows that BioPerl has more active developers than the two other projects. Figure 2C shows average number of monthly commits (commit frequency) for BioPerl is 92.62. However, these numbers for BioJava and BioRuby are 42.22 and 17.58 respectively. Comment density result presented in Figure 2D shows an average comment density of 0.42 for BioPerl, 0.35 for BioJava, and 0.30 for BioRuby. As discussed previously, the higher percentage of comment density can be interpreted as an indicator of better quality and maintainability of the project [13].
4. Limitation and Future Work

The data obtained for this project was from a crawler that counts physical lines and does not examine its contents. Thus, in terms of actual comment density, our results may be misleading. For instance, large number of comments in open source software are auto-generated, or comments may refer to the license that is used within the project [2][13]. We do not feel that this is a major problem for our study, as all three presented languages are open source projects and this is true for all three languages and does not affect the result of comparison.

As an extension of this current study, we plan to develop prediction model for popularity of language. Additionally, we plan to collect more quantitative metrics from software repositories, and apply them for qualitative analysis of these three languages. Also of interest are to analyze and predict the reliability of each open source language through the study of defect density and failure probability.

Table 2. Summary results from Figures 2A to 2D

<table>
<thead>
<tr>
<th>Qualitative Parameters</th>
<th>Quantitative Metric</th>
<th>BioPerl</th>
<th>BioJava</th>
<th>BioRuby</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintainability</td>
<td>Comment Density</td>
<td>0.42</td>
<td>0.35</td>
<td>0.3</td>
</tr>
<tr>
<td>Popularity</td>
<td>Average Number of Committers per month</td>
<td>9.73</td>
<td>4.16</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td>Average Commit Frequency per month</td>
<td>92.62</td>
<td>42.22</td>
<td>17.58</td>
</tr>
<tr>
<td></td>
<td>Total LOC in July 2011</td>
<td>717,749</td>
<td>407,158</td>
<td>109,478</td>
</tr>
</tbody>
</table>

5. Summary and Conclusion

This paper is an effort to compare software repository metrics to understand maintainability and popularity of three different languages used in bioinformatics domain. Four different repository metrics from three languages were selected and compared with one another. Our results show that BioPerl is the dominant bioinformatics language. With a largest LOC, one can find more functionality in BioPerl than in the two other languages. A higher number of committers and commit frequency in BioPerl reflected an active community of developers. This is indicative of open source project that can grow with advancement in technologies and new discoveries in the bioinformatics field. The highest percentage of comment density seen in BioPerl compared to BioJava and BioRuby is also a good indicator of high quality code that its developers write. These results can be helpful for developers who want to choose an appropriate language for the development of bioinformatics applications.

References