**b-Src: A Source-Code Search Engine for Bioinformatics Open-Source Packages**

Mitsuteru Nakao  
nakao-mitsuteru@aist.go.jp

Paul Horton  
horton-p@aist.go.jp

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

**Keywords:**  open-source software, source-code, programming, search engine

1 Introduction

Computer programming is one of the most fundamental skills for bioinformatics. In varying degrees, researchers who focus on bioinformatics invest time not only in typing computer programs but also finding out the usage of programming languages and libraries.

*b-Src*[1] is a web-based source-code search engine for bioinformatics packages powered by gonzui[2], a source-code search engine. Our goal is to help bioinformatists and biological researchers develop software effectively by enabling them to quickly find appropriate libraries and implementations of analysis schemes and algorithms in open-source packages.

2 System Overview

2.1 Search Engine

*b-Src* contains over 160 open-source bioinformatics and bioinformatics related packages. *b-Src* provides keyword search for the packages (Fig. 1) powered by gonzui[2]. Gonzui provides lexical parsing for various computer languages and contents indexing based on the BerkleyDB. The indexed packages consist of following languages (or file formats): Text, C/C++, Binary, Java, Perl, Ruby, Python, shell script, Automake, CSS, Objective Caml, Emacs Lisp, JavaScript, RPM SPEC and Autoconf. *b-Src* provides various keyword and phrase search: (1) limited within a specific target (lexical), format (computer programming language) or licence (GPL etc ...) and (2) limited within a specific package.

2.2 Source-code Browser

*b-Src* is also a source-code browser with syntax highlighting and useful source-code specific information. Files in a package are listed along their directory hierarchy. Users have easy access to files in the target package. Users can access appropriate files along embedded hyperlinks at lexical parsed words (class, function and variable names). Syntax based highlighting makes the source-code easy to read.

2.3 Servers

*b-Src* web server[1] is running on a Debina GNU/Linux box. The server PC contains 2GB memory and 2.80GHz Intel Pentium 4 CPU. Usually the gonzui HTTP server process occupies hundreds of mega bytes of memory.
3 Discussions

Over 160 packages frequently used for bioinformatics tasks are indexed in the *b-Src* server, which include so-called Bio\* packages[3], NCBI Tools, statistics related perl modules etc... A complete list of packages is available at http://b-src.cbrc.jp/markup. It consists of 58K files, 27M lines and 3M indexed keys. Most files are in plain text. C/C++ source files are the most frequent programming language. A complete report about index statistics is available at http://b-src.cbrc.jp/stat.

*b-Src* is a domain specific service strategy. This strategy provides search results with small false positives relative to general web search engines.

References

