The Bioscanners Project at Sourceforge

Detlef Groth (1), Stefan Müller (1), Joachim Selbig (1)

(1): Potsdam University, Bioinformatics Group, c/o Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany

Although biologists and bioinformaticians are permanently challenged with parsing and analysing output of various biological tools, often utilising different formats, a standard method for solving such tasks has not been found, yet. To overcome the problems of commonly used standalone applications (difficult data integration), and Bio-frameworks (complex programming interface, slow data parsing), we generated the Bioscanners project. The usage of scanner generators in application coding ensures easy programming, little influence of personal programming styles, a small code base, easy maintenance, and very high processing speed.

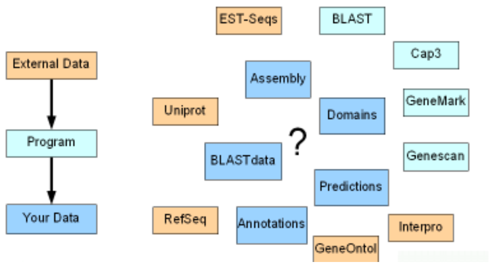


Figure: The data integration problem: How to look at all those different formats? How to make relations between all those data?

Scanner Advantages

- Easier to program than traditional parsers
- Standalone programs
- One file - one application
- No installation, just download and execute
- Almost no library dependencies (just libC)

Besides easy and fast data scanning, the storage of the scanning result in a format, that can be efficiently used in downstream processing and to selectively extract information, is important for the user. For that reason the output of our scanners is standard database code, suitable to be used for SQL compliant databases like PostgreSQL, MySQL or SQLite. That data integration platform ensures the best possible performance in data exploration, especially in a relational context, as well as in maintaining data integrity.

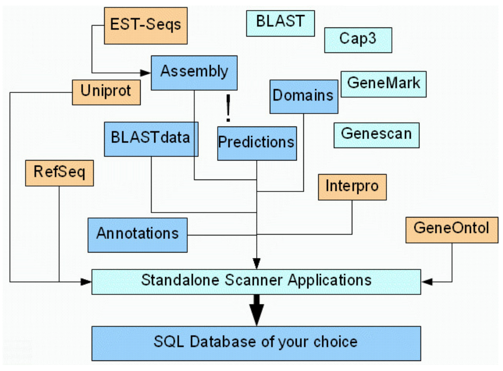


Figure: The data integration solution: Write small programs for data scanning with SQL output. That output can be directly piped in a database.

Database advantages

- All data in one place (even one file with SQLite)
- No API required, just ask your data using SQL
- Easy to learn, within minutes
- Sophisticated GUIs are available for interactive data queries and data exploration

As an example, our BLASTScanner, available from the sourceforge project page, is a single C source file which can be compiled on any modern computer platform to a small 20Kb executable not depending on any external library or runtime. The data processing time required of the scanner is shorter than the time required by the database to import the data. Even input files of several hundred MB can be translated into database code within several seconds.

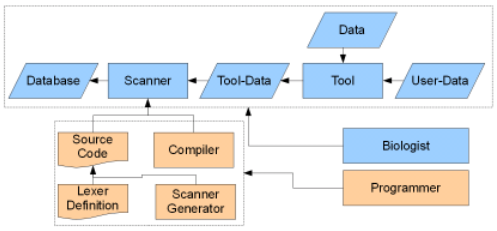


Figure: Data-Analysis pipeline using a scanner generator. Programmers and biologists work is strictly separated.

```
$ blastall -p blastn -i sample.fasta -d mydb > sample.blastn
$ BLASTScanner-Linux-x86 --infile sample.blastn \
```

```
       --prefix saml | sqlite3 sample.db3
$ sqlite3 sample.db3 "select count(*) from saml_hits where score > 500"
3428
```

Figure: Invocation of BLASTScanner with a redirect to the database application of sqlite3.

Table 1. Scanners, Compilers and WC binary sizes

We compared different programming languages and scanner generators concering their speed, memory and binary sizes (+++ best, ++ very good, + good, - bad). wc-sizes are size of an word counter implementation in kB.

scanner	lang.	reference	compiler	wc-sizes	speed/mem
Flex	C	Paxson (1995)	GNU gcc 4.1.2	16 / 489	++ / ++
Flexpp	C++	Paxson (1995)	GNU g++ 4.1.2	21 / 1002	- / +
RE2C	C	Bumbulis and Cowan (1993)	GNU gcc 4.1.2	7 / 7	+++ / ++
TPLex	Pascal	Graef (1998)	Free Pascal fpc 2.0.4	109 / 109	+ / +++
JFlex	Java	Klein (2008)	SUN javac 1.6.0	7 / NA	+ / -
GPLEX	C#	Gough (2008)	Mono gmcs 2.0.1	13 / 6525	- / -

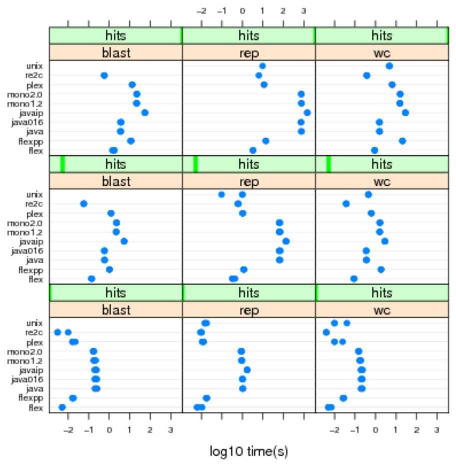
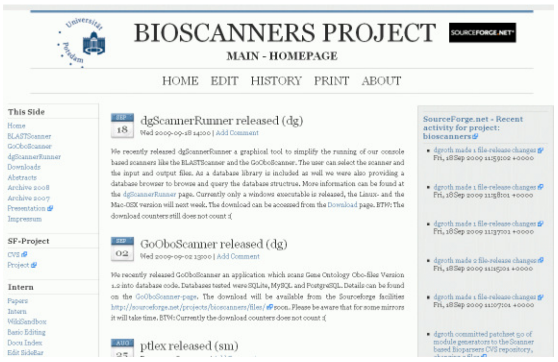


Figure: Time usage for different scanners and tasks. wc=word counter, rep=simple character replacer, blast=simple blast scanner. Input were BLAST files with 1 (left), 100 (center) and 1000 (right) hits. The latter file was around 40Mb large.

Conclusions

Scanner applications can serve as a platform for the development of standardized bioinformatics related tools. The source code for some proof of prinicple implementations is freely available at the sourceforge project page under an open source BSD-license. Programmers, who would like to participate in the development are highly welcome.



BIOSCANNERS PROJECT

MAIN - HOMEPAGE

HOME EDIT HISTORY PRINT ABOUT

This Side
Home
BLASTScanner
GoOboScanner
dgScannerRunner
Downloads
Abstracts
Archive 2008
Archive 2007
Presentation
Impressum

SF-Project
CVS
Project

Intern
Papers
Intern
WikiSandbox
Basic Editing
Docu Index
Edit SideBar

SEP 18 dgScannerRunner released (dg)
Wed 2009-09-18 14:00 | Add Comment

We recently released dgScannerRunner a graphical tool to simplify the running of our console based scanners like the BLASTScanner and the GoOboScanner. The user can select the scanner and the input and output files. As a database library is included as well we were also providing a database browser to browse and query the database structrue. More information can be found at the dgScannerRunner page. Currently only a windows executable is released, the Linux- and the Mac-OSX version will next week. The download can be accessed from the Download page. BTW: The download counters still does not count :(

SEP 02 GoOboScanner released (dg)
Wed 2009-09-02 13:00 | Add Comment

We recently released GoOboScanner an application which scans Gene Ontology Obo-files Version 1.2 into database code. Databases tested were SQLite, MySQL and PostgreSQL. Details can be found on the GoOboScanner-page. The download will be available from the Sourceforge facilities http://sourceforge.net/projects/bioscanners/files/ soon. Please be aware that for some mirrors it will take time. BTW: Currently the download counters does not count :(

AUG 25 ptlex released (sm)

SourceForge.net - Recent activity for project: bioscanners
- dgroth made 1 file-release changes
  Fri, 18Sep 2009 11:59:02 +0000
- dgroth made 1 file-release changes
  Fri, 18Sep 2009 11:58:01 +0000
- dgroth made 1 file-release changes
  Fri, 18Sep 2009 11:37:01 +0000
- dgroth made 2 file-release changes
  Fri, 18Sep 2009 11:15:01 +0000
- dgroth made 1 file-release changes
  Fri, 18Sep 2009 11:07:01 +0000
- dgroth committed patchset 50 of module generators to the Scanner based Bioparsers CVS repository,

Figure: The website of the bioscanners project at http://bioscanners.sf.net

- RE2C generated scanners are single C-code files
- Can be compiled with any modern C-compiler
- No external dependencies, run just standalone
- 20-80kb binaries (Linux, Win32, OSX, Solaris, OSF-1, ...)
- Example: on a normal modern desktop PC, BLASTScanner translates 50MB BLAST-file in one second into database code
- The database actually needs longer to import the code
- Scanners are magnitudes faster than the parser of the Bio*-approaches
- Other scanners for BLAST-M8 files, Gene Ontology Obo-files, gene predictor files are available on the project page as well
- Vizit an join us at: http://bioscanners.sf.net
- **Contact:** dgroth@uni-potsdam.de