

# BioParser: Easier Mining of Online Bioinformatics Databases

Carol Barner, Intern

John Pearson, BioParser Creator

Dr. Rosemary Renaut, Internship Advisor

Dr. Eric Kostelich, Committee Member

Dr. Martin Wojciechowski, Committee Member

# Motivation

- Many online bioinformatics databases
- Easy to mine individual records
- Hard to mine thousands of records
- Text file downloads
  - Huge
  - Frequent data updates
  - Frequent file format changes

# BioParser

- John Pearson of TGen
- Parse and convert the files to objects
- Easy, fast searching
- All fields in many online databases
- Reliable, commercial-quality support

# Project Goals

- Easy, consistent, fast data mining
- An internal TGen research tool
- A commercial product

# Procedure for Users

- Download the latest BioParser object version of a desired database.
- Insert desired fields in sample code.
- Specify additional selection criteria.
- Execute to produce a file or report.

# BioParser Features

- Object-oriented Perl
- Parser
- Systematic Organization
- Testing Harness
- Auto-generated Documentation
- Concurrent Versioning System
- Unix and VPN

# Object-Oriented Perl

- objects
- regular expressions and hashes
- well-integrated with Unix
- chosen parser is written in Perl
- popular language in bioinformatics databases

# Parser

- Character by character
- Recognize data fields
- Variability
- Frequent file changes

# Parse::RecDescent

- written by Damian Conway
- powerful, flexible
- one program
- written in Perl
- syntax similar to Perl regular expressions
- free

# Parse::RecDescent Strategy

- Recursive searching
- backtracking
- syntax for what is allowed or not allowed
- may place actions after each definition

```
#!/usr/bin/perl -w
use Parse::RecDescent;
my $grammar = q{
sentence:      <skip:[' \t,']*>
                subject
                predicate
                '!'
                {print "Valid sentence.\n"}

subject:       definite_article(?)
                adjective(s?)
                noun

definite_article: 'a'|'A'|'The'|'the'

adjective:    'slow'|'quick'|
                'lazy'|'brown'

noun:         'cat'|'dog'|'fox'
```

```
predicate:    verb
                adverb(s?)
                direct_object(?)

verb:         'sprinted'|'jumped'

adverb:       'over'|'under'

direct_object: definite_article(?)
                adjective(s?)
                noun

};
my $parser = new
                Parse::RecDescent($grammar);
undef $/;
my $text = "The quick, brown fox
            jumped over the lazy dog.";
$parser->sentence($text);
```

# BioParser Parsing

- position on line or in file
- labels on fields
- indentation spaces
- stores recognized items into a hash
- returns the hash address

# Systematic Organization

- same 4 filenames for each database:  
FileParser.pm, Record.pm,  
RDParser.pm, SerialDatabase.pm
- each database has own directory
- directories for scripts, data, tests
- object hierarchy

# Testing Harness

- Standardized testing with Make Test
- Testing Procedures:
  - Perl compilation
  - basic object testing program = harness
  - bpr\_parse.pl parses a data file

# Auto-Generated Documentation

- Perl pod text file documentation
- generated from special comments
- `bpr_pod2html.pl` extends pod to html

# CVS

- Concurrent Versioning System
- multiple programmers and versions
  - keeps track of versions
  - notifies when version conflicts
  - attempts to help reconcile conflicts
- secure file storage

# Unix and VPN

- Unix can handle thousands of files in one directory
- VPN provides secure remote access

# BioParser Results

- Previously completed
  - LocusLink
  - OMIM
  - Unigene
  - TIGR Resourcer
  - UniProt
- Partially completed previously
  - PDB
  - GenBank
- This internship project
  - Homologene
  - dbEST

# Homologene

41	9606	675	BRCA2	4502451		NP_000050.1
41	9598	452526	LOC452526		55639693	XP_509619.1
41	9615	474180	BRCA2	57104978		XP_543141.1
41	10090	12190	Brca2	6857765		NP_033895.1
41	10116	497682	LOC497682		62658293	XP_579511.1
41	9031	374139	BRCA2	45383586		NP_989607.1
41	7165	1270068		1270068	31199573	XP_308734.1



# A dbEST Sample Record



||

## IDENTIFIERS

dbEST Id: 30612150  
 EST name: BW974700  
 GenBank Acc: BW974700  
 GenBank gi: 71959383

## CLONE INFO

Clone Id: PBL010097E04 (5')  
 DNA type: cDNA

## PRIMERS

PolyA Tail: Unknown

## SEQUENCE

GAGACTCTTGAGAAGGCAGCTACGGAGGCTGCAGAGGTCTGGCAGGCCATGGAGGAGCCC  
 CCTTTGCGAGAGGAGGAGGAGGAAGGGGACGAGGCGGGGCCCGAGGGGGCTCTGGGCAAG  
 AGCCCCTTCCAGCTGACAGCCGAAGACGTATATGACATCTCTTACGTGATGGGCCGAGAG  
 CTGATGGCCCTGGGCAGCGACCCCCGGGTGACACAGCTGCAGTTCAAGATCGTCCGTGTT  
 CTGGAGATGCTGGAGACGCTGGTGAATGAGGGCAACTTGACGGTGGAGGAGCTGAGAATG  
 GAGCGGGACAACCTCAGGACGGAGGTGGAGGGCTGCGGAGAGAGGGCTCCGCGGCCGGC  
 GGAGAGGTGAACCTGGGACCAGACAAAATGGTGGTTGACCTGACAGATCCCAACCGACCA  
 CGCTTTACTCTGCAGGAGCTGAGGGATGTGCTACAGGAGCGCAACAACTCAAGTCGCAG  
 CTGCTGGTGGCACAGGAGGAGCTGCAGTGCTATAAGAGTGGCCTGATTCCACCAAGAGAA  
 GGCCAGGAGGAAGAAGAGAAAAAGATACTCTGGTTGCTCGGGCCAACAATGCCAGGAGT  
 AACAAGGAGGAGAAGACAATCATAAGGAAGCTGTTCTCTTTTCGGATCAGGGAAGCAGACA  
 TAGATCTGAGGCCACGACTAAATTCTCAGACTCAGAAAACAGCTCACAAAGACAACCTTCC  
 AAAATCATCTCTCAGTGCCACGCGTACCCACTGCACATGCTGCTTTGTTTCTCCTCAAAG  
 CTGTCTGAGGAGGAAGGGGAAACGTTTTCTCCCTAGCTGCAGAACTGGACACCCTTGAAG  
 GCTGGGCCAGAGCAGA

Entry Created: Aug 8 2005

Last Updated: Aug 8 2005



## COMMENTS

EST project with full-length enriched cDNA libraries carried out in Animal Genome Research Program (Japan) by National Institute of Agrobiological Sciences and STAFF-Institute Single pass sequencing of clones derived from oligo-capped cDNA library

Vector sequences were eliminated by RepeatMasker version 2002/07/13 and crossmatch version 0.990319

Low quality bases were trimmed based on the quality values

## LIBRARY

dbEST lib id: 15103

Lib Name: full-length enriched swine cDNA library, adult peripheral blood mononuclear cell

Organism: *Sus scrofa*

Tissue type: peripheral blood mononuclear cell

Develop. stage: adult

## SUBMITTER

Name: Hirohide Uenishi

Lab: Animal Genome Laboratory, Genome Research Department

Institution: National Institute of Agrobiological Sciences

Address: 2 Ikenodai, Tsukuba, Ibaraki 305-8602, Japan

Tel: +81-29-838-8627

Fax: +81-29-838-8627

E-mail: huenishi@affrc.go.jp

## CITATIONS

Medline UID: 14681463

Title: PEDE (Pig EST Data Explorer): construction of a database for ESTs derived from porcine full-length cDNA libraries

Authors: Uenishi,H., Eguchi,T., Suzuki,K., Sawazaki,T., Toki,D., Shinkai,H., Okumura,N., Hamasima,N., Awata,T.

Citation: Nucleic Acids Res. 32 (1): D484-D488 2004

MAP DATA

||

# Conclusions and Future

- a valuable scientific data mining tool
- Thirteen databases planned
- finish in 2006
- commercially marketable
  - Licenses for downloading object files
  - Revenue for internal TGen projects

# Existing Alternatives

- Parsers written for 1 or 2 databases
- Extraction programs such as BioPerl
  - many databases
  - extract only common fields
- Volunteer support, slow speed
- Difficult commercial programs
- Write a custom program

# References

- 1) GenBank: <http://www.ncbi.nlm.nih.gov/Genbank/>
- 2) Swiss-Prot: <http://www.ebi.ac.uk/swissprot/>
- 3) BioParser: <http://bioinformatics.tgen.org/brunit/software/bioparser/>
- 4) Barner, Carol, Gholba, Sumeda, Goldberg, Loretta, Gupta, Pankaj, and Revollo, Julio, *Venture Feasibility Study for BioParser*, prepared for Professor Bradford Kirkman-Liff, HSM 591, December, 2005.
- 5) BioPerl: <http://www.bioperl.org>
- 6) Perl: <http://www.perl.com/>
- 7) Parse::RecDescent: <http://search.cpan.org/~dconway/Parse-RecDescent-1.94/lib/Parse/RecDescent.pod>
- 8) Parse::RecDescent tutorial: <http://search.cpan.org/src/DCONWAY/Parse-RecDescent-1.94/tutorial/tutorial.html>
- 9) CVS : [http://ximbiot.com/cvs/wiki/index.php?title=Main\\_Page](http://ximbiot.com/cvs/wiki/index.php?title=Main_Page)
- 10) Homologene : <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene>
- 11) dbEST: <http://www.ncbi.nlm.nih.gov/dbEST/>
- 12) Perl pod : <http://perldoc.perl.org/perlpod.html>
- 13) UNIX: <http://www.unix.org/>
- 14) VPN: <http://www.vpnlabs.com/>