G-language

Genome Analysis Environment

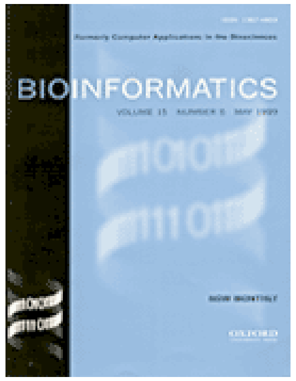
Kazuharu Arakawa
Institute for Advanced Biosciences
Keio University

G-language Genome Analysis Environment

2001: Institute for Advanced Biosciences, Keio University

Make developers' lives easier for bioinformatics
-Perl library
-interactive shell
-graphical user interface

Open source：GPL（partly LGPL)



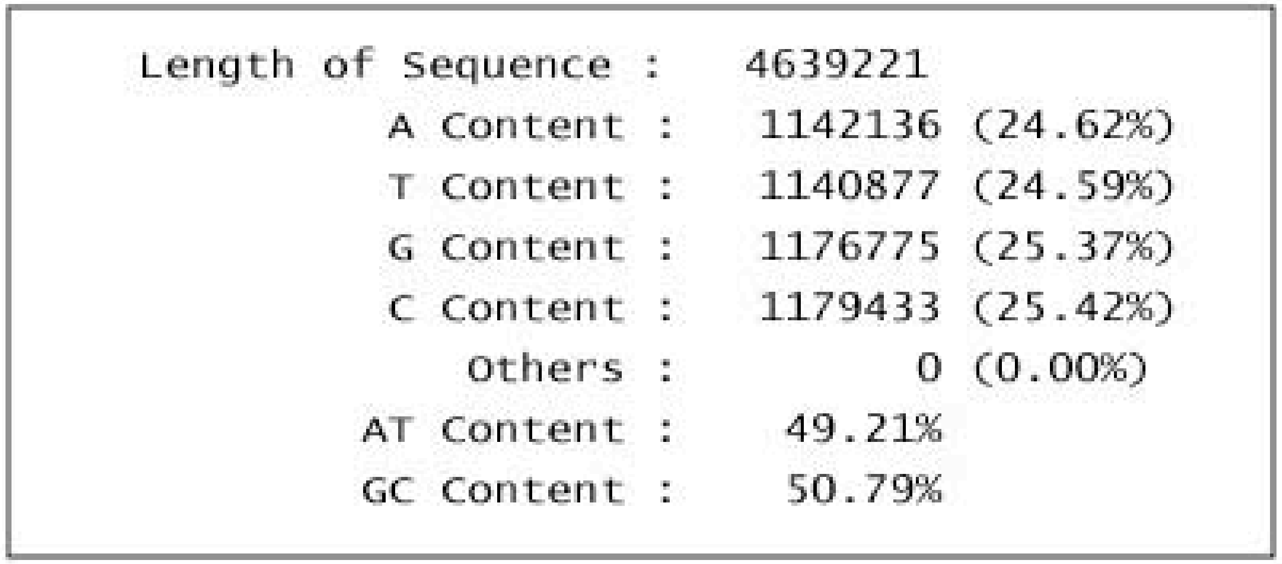
G-language Genome Analysis Environment:
a workbench for nucleotide sequence data mining
K. Arakawa, et al.
Bioinformatics 2003, 19(2):305–306

For programmers:

GC skew Analysis

```
use G;
$gb = new G("ecoli.gbk");
genomic_skew($gb);
```

stdout

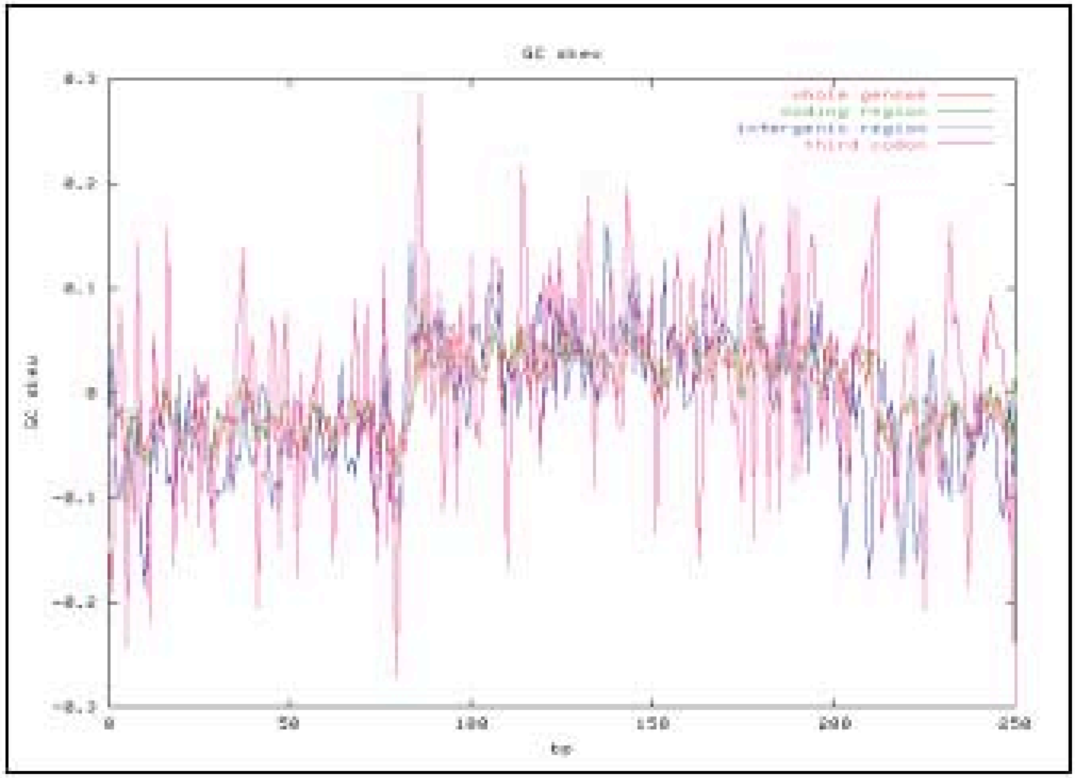


```
Length of Sequence :    4639221
         A Content :    1142136 (24.62%)
         T Content :    1140877 (24.59%)
         G Content :    1176775 (25.37%)
         C Content :    1179433 (25.42%)
            Others :          0 (0.00%)
        AT Content :     49.21%
        GC Content :     50.79%
```

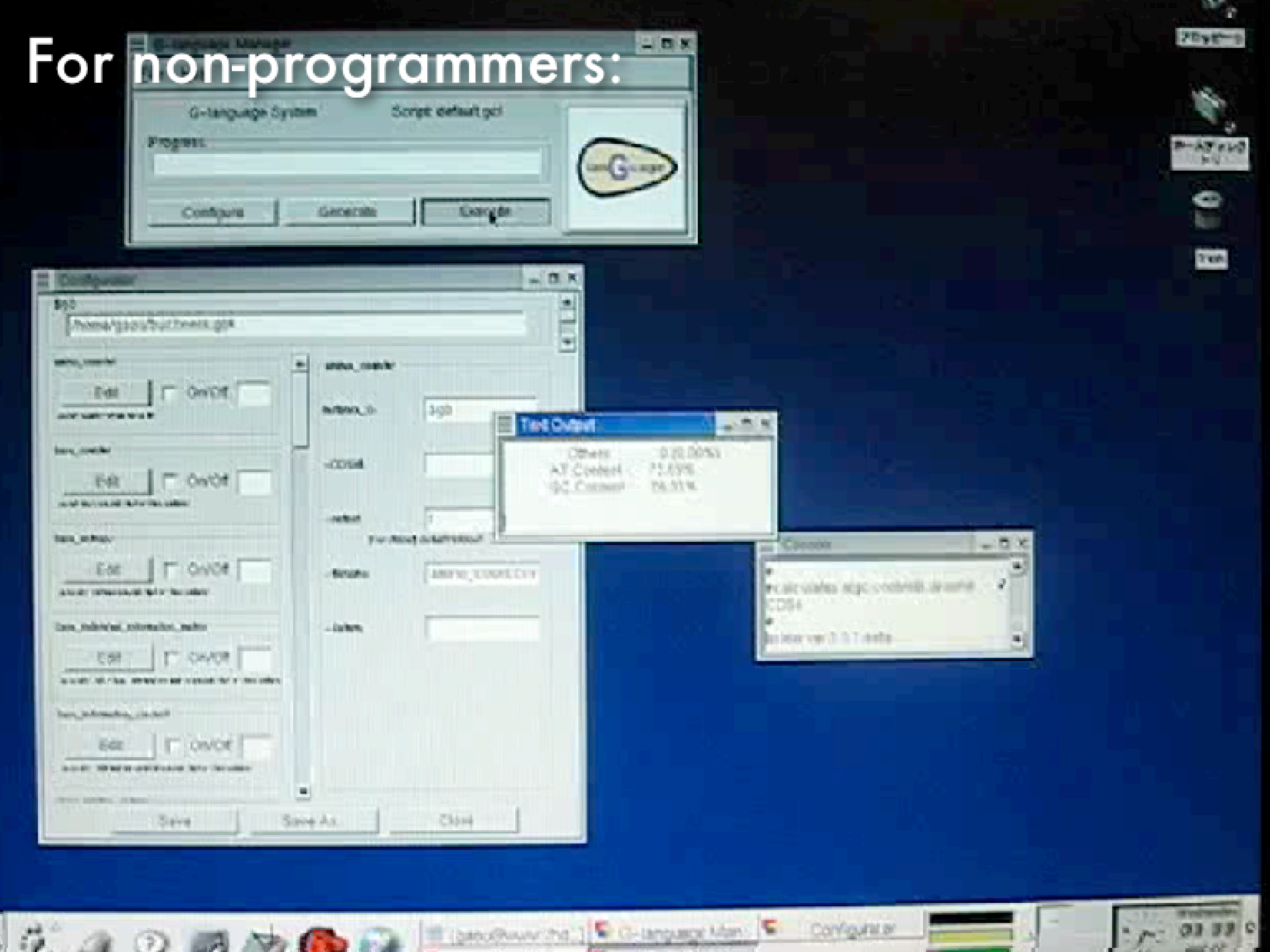
To perform GC skew analysis
for all genes:

```
foreach ($gb->cds()){
   gcskew($gb->get_geneseq($_));
}
```

pop-up window



For non-programmers:



Why not Bioperl?

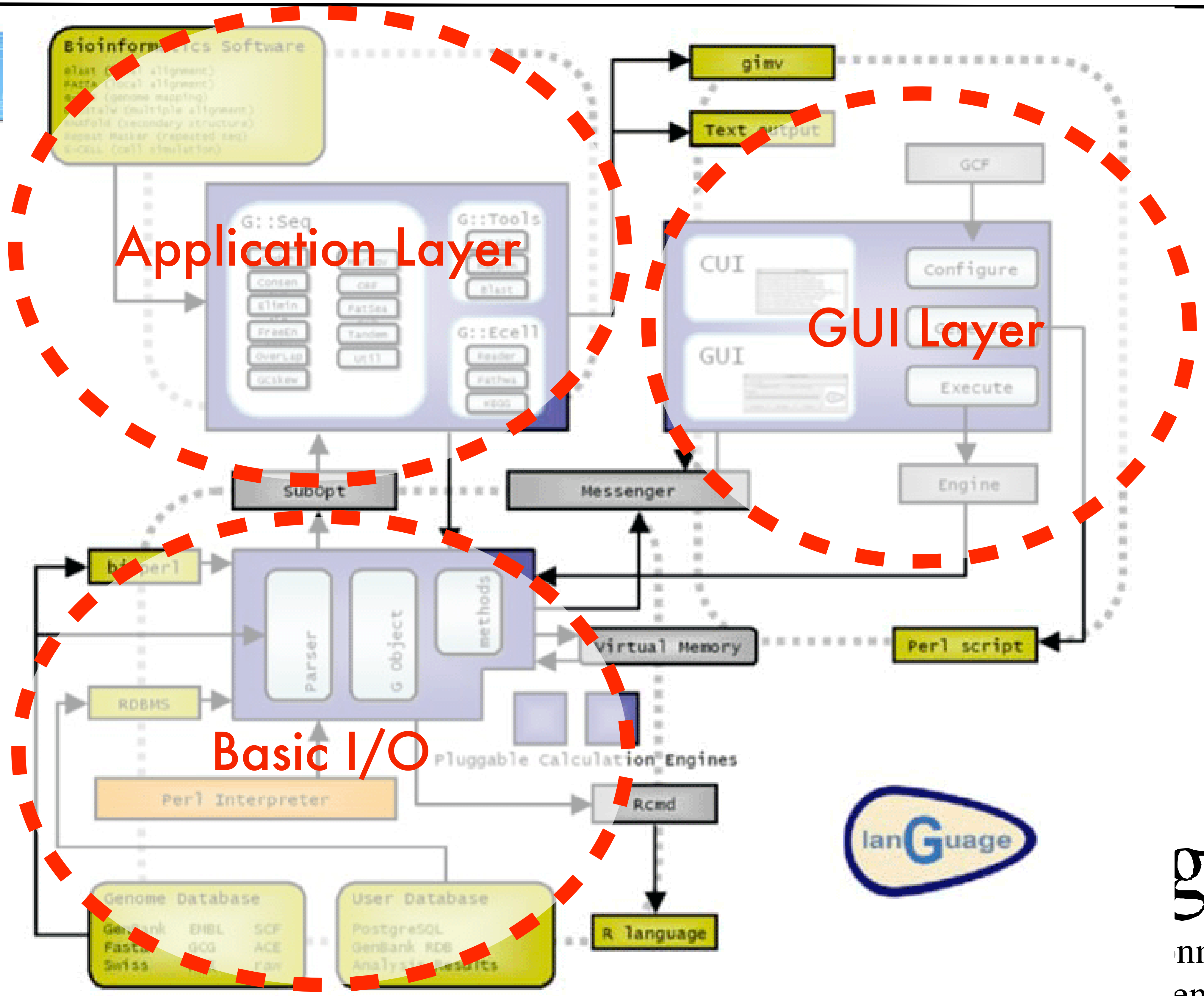
-Focus on development environment

-Focus on accessibility and visualization
certain framework is necessary for GUI conversion

-Target on systems biology (application is discussed on poster)

-but we have no purpose in “re-invention of the wheel”
we actually use bioperl in G-language GAE

G-language 2 se
Genome Analysis Environment
Simulation Environment



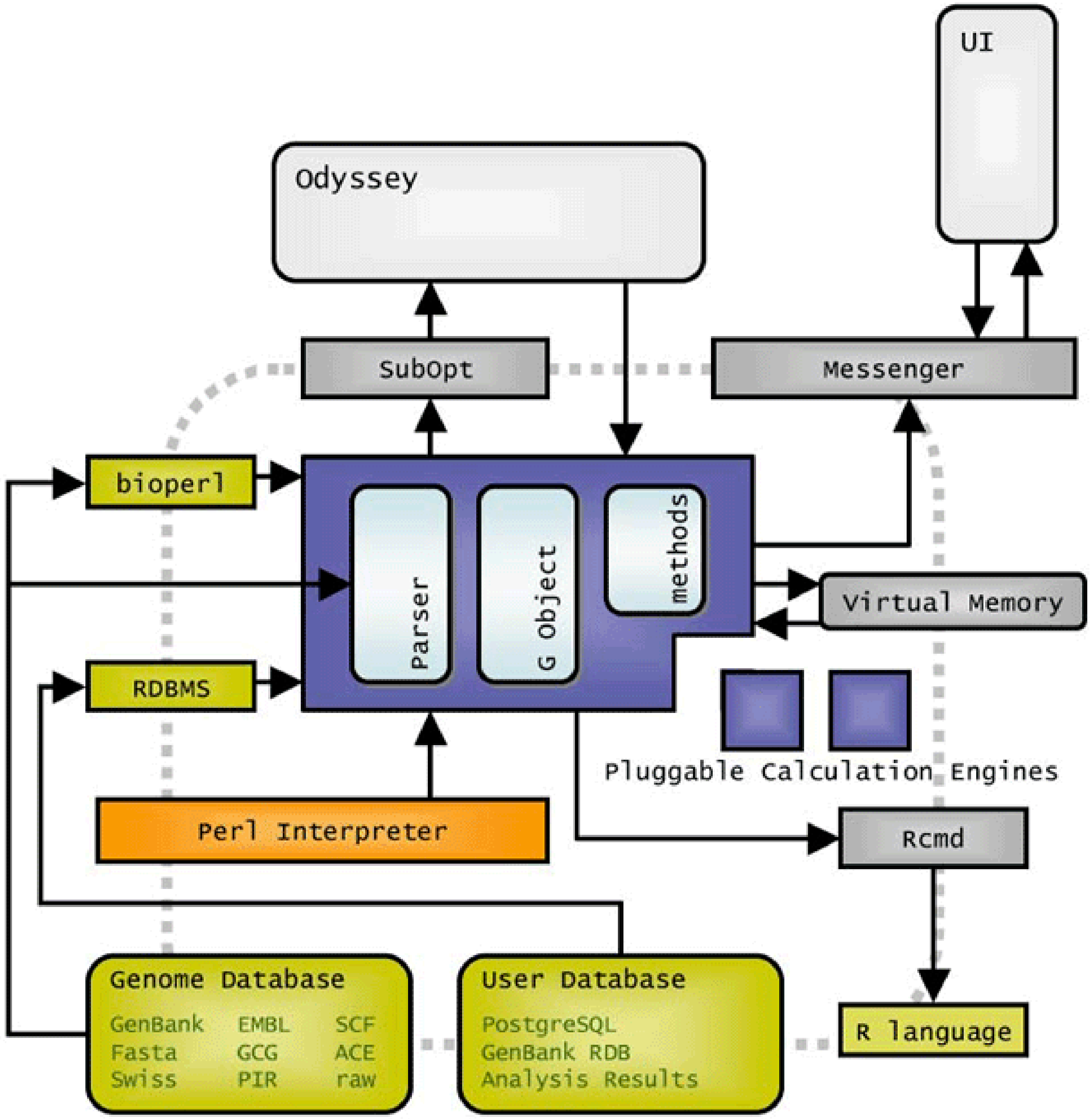
se

ge

nment

ent

Database I/O



- I/O via BioPerl

GenBank, Fasta, EMBL, Swiss, PIR, SCF, GCG, Ace, raw, Qual, Phd, BSML

- Original I/O (>2.5 times faster)

GenBank, Fasta, EMBL, Swiss

- Systems Biology

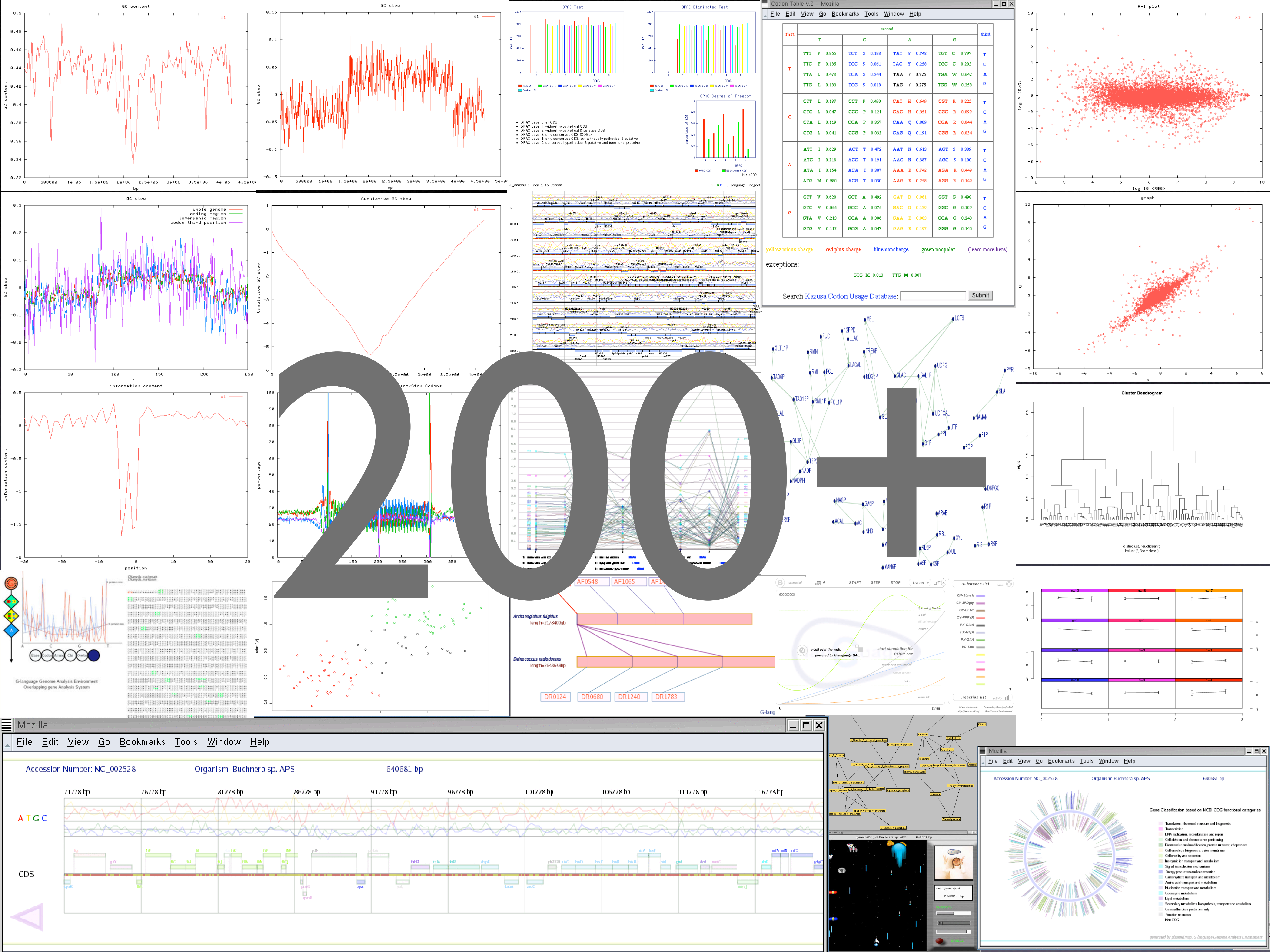
ptt, eri, EML, SBML, KEGG, Brenda, WIT, BioCyc

G-language 2 se

Genome Analysis Environment
Simulation Environment

Bioperl compatibility

```
$gb = new G(”genome.embl”); #load EMBL file to G instance
$bp = $gb->bioperl();       #create bioperl instance from G
$gb = new G($bp);           #create G instance from bioperl
gcskew($gb);                #call G method with G instance
gcskew($bp);                #call G method with bioperl instanc
```



200+

Codon Table v.2 - Mozilla

first	second T	second C	second A	second G	third
T	TTT F 0.865	TCT S 0.188	TAT Y 0.742	TGT C 0.797	T
	TTC F 0.135	TCC S 0.061	TAC Y 0.258	TGC C 0.203	C
	TTA L 0.473	TCA S 0.244	TAA / 0.725	TGA W 0.642	A
	TTG L 0.133	TCG S 0.018	TAG / 0.275	TGG W 0.358	G
C	CTT L 0.187	CCT P 0.490	CAT H 0.649	CGT R 0.225	T
	CTC L 0.047	CCC P 0.121	CAC H 0.351	CGC R 0.099	C
	CTA L 0.119	CCA P 0.357	CAA Q 0.809	CGA R 0.044	A
	CTG L 0.041	CCG P 0.032	CAG Q 0.191	CGG R 0.034	G
A	ATT I 0.629	ACT T 0.472	AAT N 0.613	AGT S 0.389	T
	ATC I 0.218	ACC T 0.191	AAC N 0.387	AGC S 0.100	C
	ATA I 0.154	ACA T 0.307	AAA K 0.742	AGA R 0.449	A
	ATG M 0.980	ACG T 0.030	AAG K 0.258	AGG R 0.149	G
G	GTT V 0.620	GCT A 0.492	GAT D 0.861	GGT G 0.498	T
	GTC V 0.055	GCC A 0.075	GAC D 0.139	GGC G 0.109	C
	GTA V 0.213	GCA A 0.386	GAA E 0.803	GGA G 0.248	A
	GTG V 0.112	GCG A 0.047	GAG E 0.197	GGG G 0.146	G

yellow minus charge red plus charge blue noncharge green nonpolar (learn more here)

exceptions:

GTG M 0.013 TTG M 0.007

Search Kazusa Codon Usage Database: Submit

G::Tools::

Alignment	clustalw wrapper
Blast	blast wrapper
Cap3	cap3 wrapper
COGs	COGs database
EPCR	EPCR wrapper
Fasta	Fasta wrapper
Glimmer	Glimmer wrapper
GOA	GOA database
GPAC	GPAC system
HMMER	HMMER wrapper
KEGG_API	KEGG_API database
Literature	PubMed wrapper
Mapping	Mapping tools
PBS	PBS grid wrapper
RCluster	Clustering wrapper
Repeat	RepeatMasker wrapper
SIM4	SIM4 wrapper

G-language 2 se
Genome Analysis Environment
Simulation Environment

G::SystemsBiology::

BioLayout	Interaction/pathway viewer with biolayout
EcellReader	Ecell file I/O
Interaction	Protein Protein Interaction Analysis
KEGG	KEGG database access
Pathway	Pathway viewer
Serizawa	Ecell wrapper

G-language 2 se

Genome Analysis Environment
Simulation Environment

G::Seq::

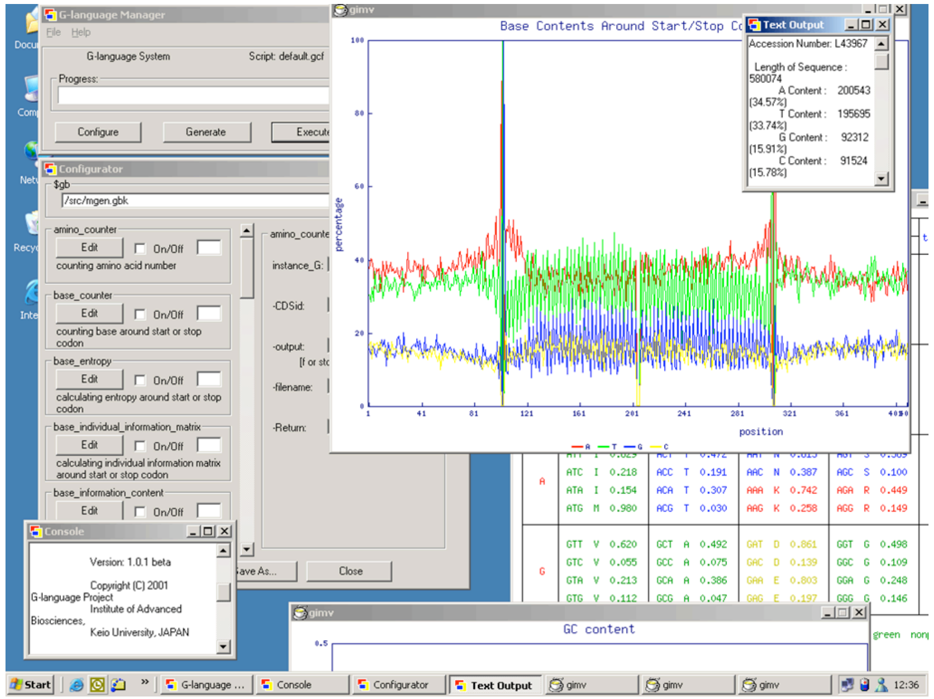
Align	alignment diffseq
AminoAcid	calc_pI amino_info peptide_mass
Codon	codon_compiler _distance_cu shannon_cu enc cbi icdi fop w_value cai phx bui aaui codon_usage
COMGA	COMGA_correlation COMGA_grapher
Consensus	base_information content, base_relative entropy base_z_value
Eliminate	valid_cds eliminate_pat
FreeEnergy	RNAfold
GCskew	gcskew, cumgcskew, find_ori_ter, leading_strand, view_cds
ImaGene	ma_normalize ma_filter ma_rfilter
Markov	markov codon_markov
Operon	set_operon
ORF	longest_ORF find_identical_gene pseudo_atg
OverLapping	overlapping_finder
PatSearch	oligomer_counter find_seq palindrome find_dnaAbox
Primitive	complement translate
Tandem	find_tandem foreach_tandem graphical LTR_search
Usage	rscu equitability cei
Util	genome_map molecular_weight plasmid_map gene_function_list

G-language 2 se
Genome Analysis Environment
Simulation Environment

~~Cross Platform GUI~~

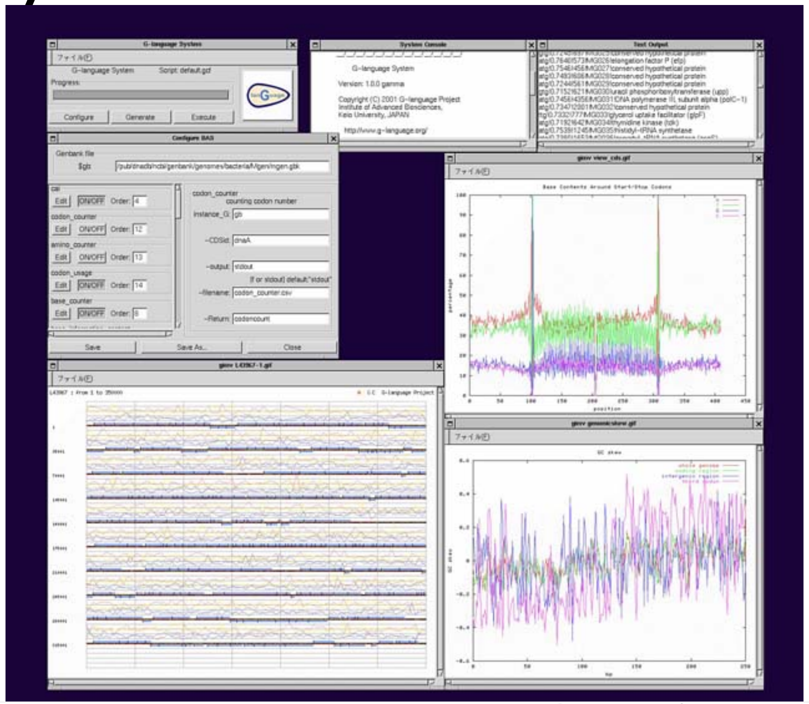
- Perl + wxWindows and CGI

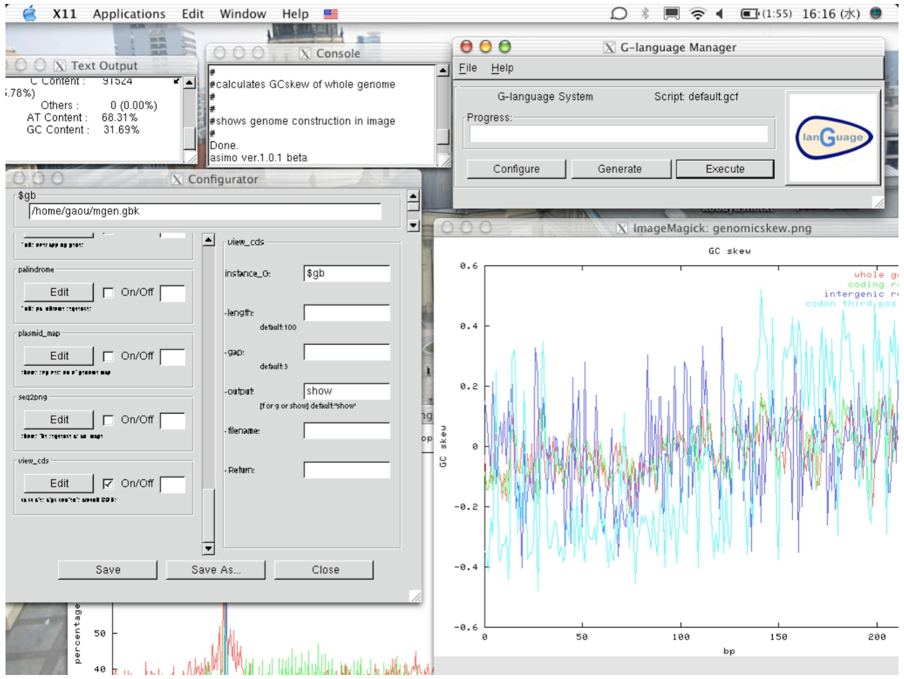
MacOS X / Windows 95,98,2000,XP / RedHat Linux



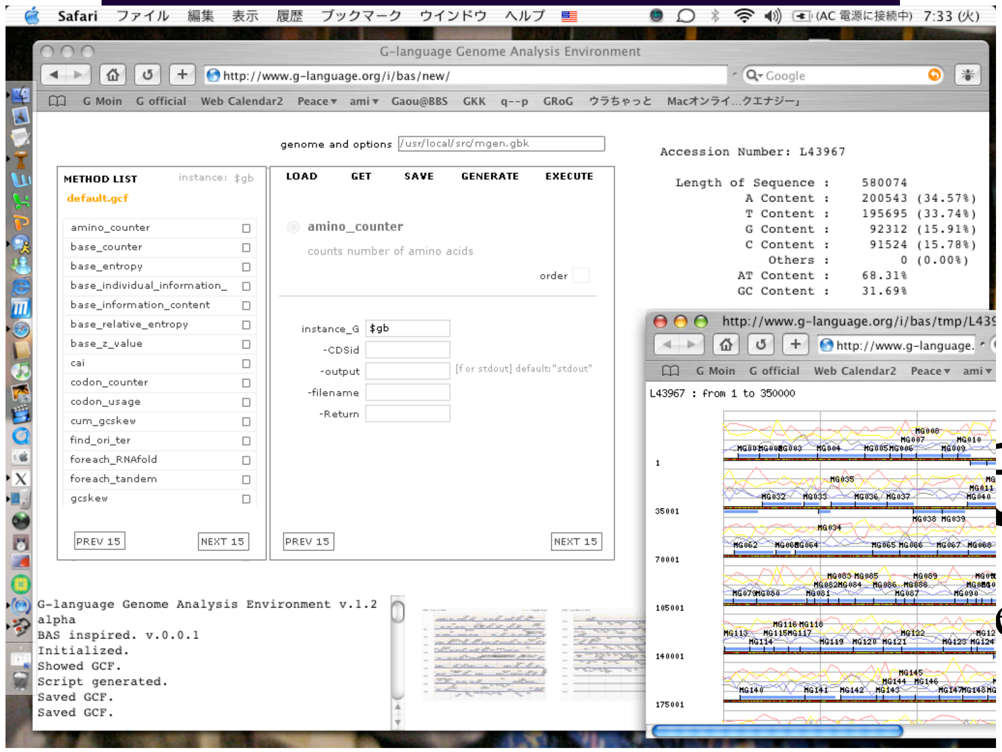
```
Accession Number: L43967

Length of Sequence :
580074
   A Content :  200543
(34.57%)
   T Content :  195695
(33.74%)
   G Content :   92312
(15.91%)
   C Content :   91524
(15.78%)
```

```
Version: 1.0.1 beta

Copyright (C) 2001
G-language Project
Institute of Advanced
Biosciences,
Keio University, JAPAN
```





```
#calculates GCskew of whole genome
#
#shows genome construction in image
#
Done.
asimo ver.1.0.1 beta
```



genome and options /usr/local/src/mgen.gbk

instance_G $gb

```
Accession Number: L43967

  Length of Sequence :    580074
           A Content :    200543 (34.57%)
           T Content :    195695 (33.74%)
           G Content :     92312 (15.91%)
           C Content :     91524 (15.78%)
              Others :         0 (0.00%)
          AT Content :    68.31%
          GC Content :    31.69%
```

```
G-language Genome Analysis Environment v.1.2
alpha
BAS inspired. v.0.0.1
Initialized.
Showed GCF.
Script generated.
Saved GCF.
Saved GCF.
```

L43967 : from 1 to 350000

se

e

ent

```
Last login: Tue Jul 15 08:03:48 2003 from 10.2.4.6
[gaou@guy ~]$ G

              _/_/_/_/_/_/_/_/_/_/_/_/_/_/

              G-language Genome Analysis Environment

               Version: 1.42
               License: GPL

               Copyright (C) 2001-2002
               G-language Project
               Institute for Advanced Biosciences,
               Keio University, JAPAN

                  http://www.g-language.org/

              _/_/_/_/_/_/_/_/_/_/_/_/_/_/

G_>$gb = new G("ecoli.gbk");


Accession Number: U00096

  Length of Sequence :    4639221
           A Content :    1142136 (24.62%)
           T Content :    1140877 (24.59%)
           G Content :    1176775 (25.37%)
           C Content :    1179433 (25.42%)
              Others :          0 (0.00%)
          AT Content :     49.21%
          GC Content :     50.79%

G_>
```

Interactive shell

Commandline interactive interpreter for short analyses and trial.

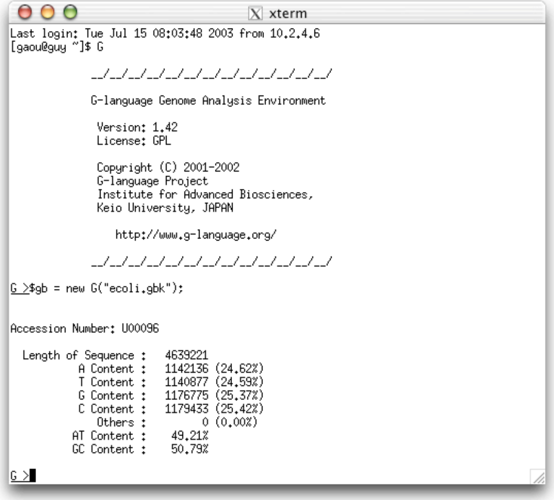
-basic shell functions
-automatic logging
-persistent memory

perl compatible.

Genome Analysis Environment
Simulation Environment

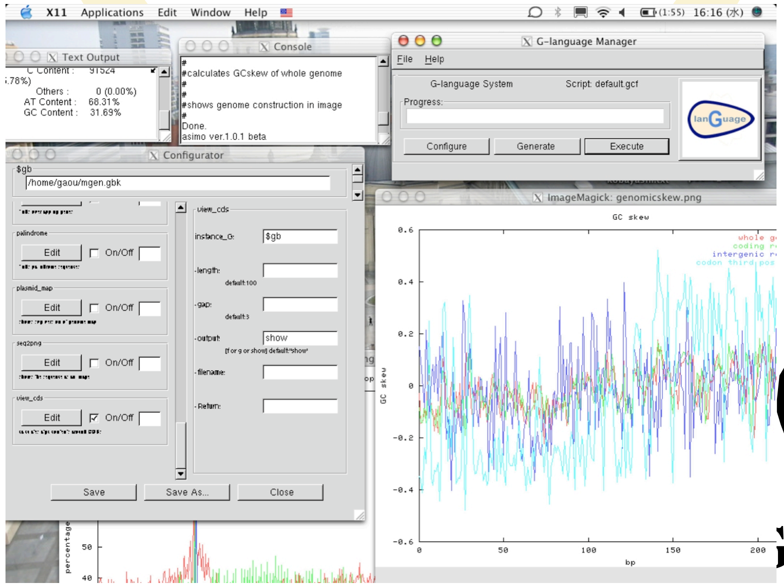
Research Cycle

generate perl script
pluggable dynamic
loader of
subroutines



```
Last login: Tue Jul 15 08:03:48 2003 from 10.2.4.6
[gaou@guy ~]$ G

            __/__/__/__/__/__/__/__/__/__/__/__/__/

            G-language Genome Analysis Environment

             Version: 1.42
             License: GPL

             Copyright (C) 2001-2002
             G-language Project
             Institute for Advanced Biosciences,
             Keio University, JAPAN

                http://www.g-language.org/

            __/__/__/__/__/__/__/__/__/__/__/__/__/

G >$gb = new G("ecoli.gbk");


Accession Number: U00096

  Length of Sequence :   4639221
           A Content :   1142136 (24.62%)
           T Content :   1140877 (24.59%)
           G Content :   1176775 (25.37%)
           C Content :   1179433 (25.42%)
              Others :         0 (0.00%)
          AT Content :    49.21%
          GC Content :    50.79%

G >
```

load perl scripts
directly converts to
GUI application
SubOpt/Messenger API



```
C Content :     91524
.78%)
    Others :      0 (0.00%)
AT Content :   68.31%
GC Content :   31.69%
```

```
#
#calculates GCskew of whole genome
#
#
#shows genome construction in image
#
Done.
asimo ver.1.0.1 beta
```

$gb: /home/gaou/mgen.gbk

palindrome ☐ On/Off

plasmid_map ☐ On/Off

seq2png ☐ On/Off

view_cds ☑ On/Off

view_cds

instance_G: $gb

-length:
default:100

-gap:
default:3

-output: show
[f or g or show] default:"show"

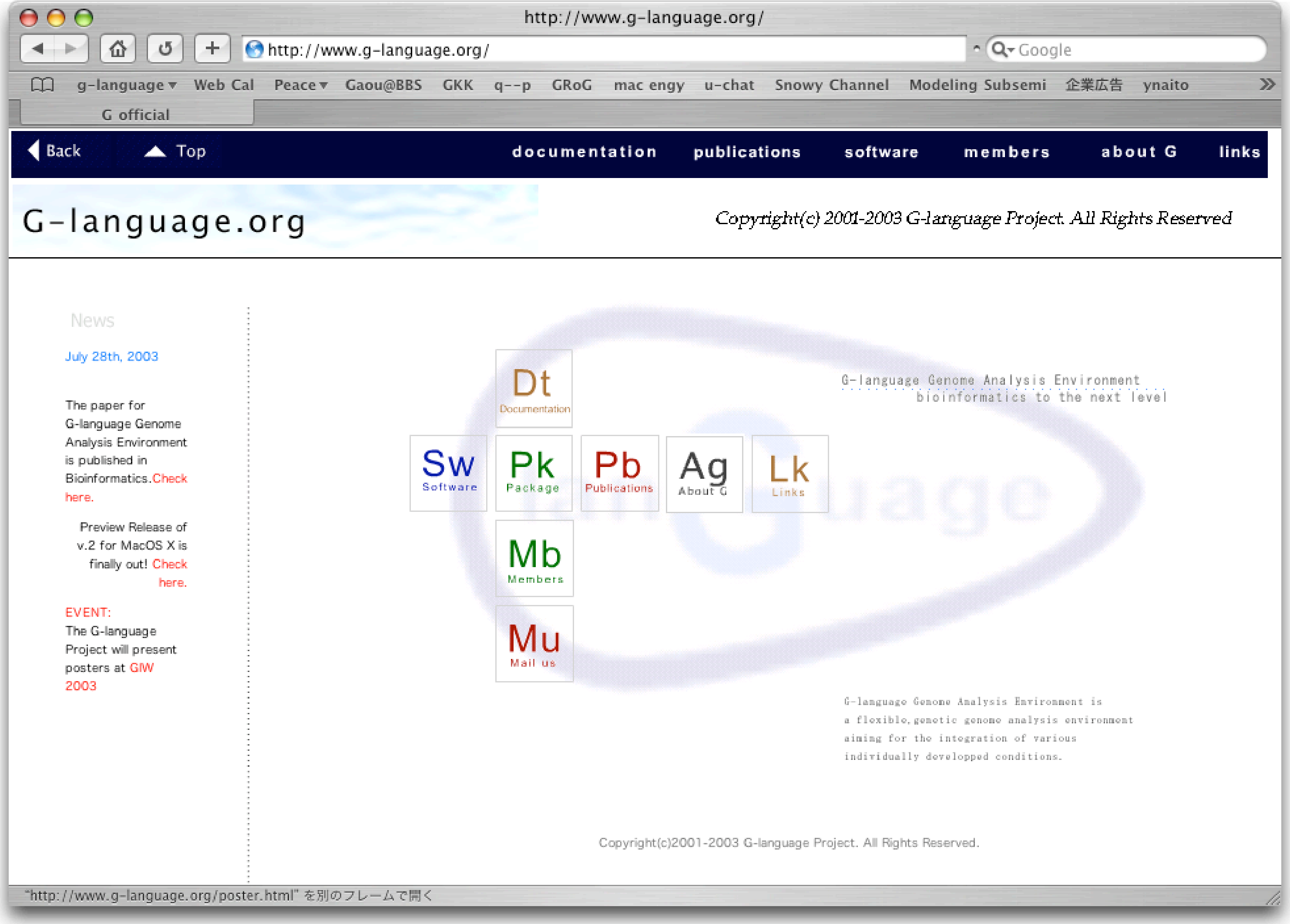
-filename:

-Return:

generate perl script

G-language ^{se}
Genome Analysis Environment
Simulation Environment

http://www.g-language.org/



Copyright(c) 2001-2003 G-language Project. All Rights Reserved

News

July 28th, 2003

The paper for G-language Genome Analysis Environment is published in Bioinformatics. Check here.

Preview Release of v.2 for MacOS X is finally out! Check here.

EVENT:
The G-language Project will present posters at GIW 2003

G-language Genome Analysis Environment
bioinformatics to the next level

G-language Genome Analysis Environment is a flexible, genetic genome analysis environment aiming for the integration of various individually developped conditions.

Copyright(c)2001-2003 G-language Project. All Rights Reserved.

"http://www.g-language.org/poster.html" を別のフレームで開く

- Software Download
- development site at bioinformatics.org
- mailing lists
- documentation and tutorials
- info@g-language.org

Acknowledgements

version 1

Bacteria Analysis System	K. Mori
cDNA Analysis System	K. Ikeda
cDNA Analysis System	T. Matuzaki
Interpreter	Y. Kobayashi
Chi Sequence Analysis System	D. Kyuma
Comparative Genome Anal. System	S. Nakamura

version 2

Bluebird DBMS	R. Hattori
Inspire Interface	Y. Yamada
Infinity Client/Server	H. Kouchi
Dynamic Loader/pl2GCF	A. Kishi
Inspire Interface	K. Shinoda
Windows port	S. Ueda
supervisor	Y. Nakayama
supervisor	M. Tomita