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**

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

To perform GC skew analysis for all genes:

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

**stdout**

```plaintext
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%
```
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
Genome Analysis Environment
Simulation Environment
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
Genome Analysis Environment
Simulation Environment
$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 instance
# G::Tools::

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td>clustalw wrapper</td>
</tr>
<tr>
<td>Blast</td>
<td>blast wrapper</td>
</tr>
<tr>
<td>Cap3</td>
<td>cap3 wrapper</td>
</tr>
<tr>
<td>COGs</td>
<td>COGs database</td>
</tr>
<tr>
<td>EPCR</td>
<td>EPCR wrapper</td>
</tr>
<tr>
<td>Fasta</td>
<td>Fasta wrapper</td>
</tr>
<tr>
<td>Glimmer</td>
<td>Glimmer wrapper</td>
</tr>
<tr>
<td>GOA</td>
<td>GOA database</td>
</tr>
<tr>
<td>GPAC</td>
<td>GPAC system</td>
</tr>
<tr>
<td>HMMER</td>
<td>HMMER wrapper</td>
</tr>
<tr>
<td>KEGG_API</td>
<td>KEGG_API database</td>
</tr>
<tr>
<td>Literature</td>
<td>PubMed wrapper</td>
</tr>
<tr>
<td>Mapping</td>
<td>Mapping tools</td>
</tr>
<tr>
<td>PBS</td>
<td>PBS grid wrapper</td>
</tr>
<tr>
<td>RCluster</td>
<td>Clustering wrapper</td>
</tr>
<tr>
<td>Repeat</td>
<td>RepeatMasker wrapper</td>
</tr>
<tr>
<td>SIM4</td>
<td>SIM4 wrapper</td>
</tr>
</tbody>
</table>

# G-language se

Genome Analysis Environment
Simulation Environment
<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioLayout</td>
<td>Interaction/pathway viewer with biolayout</td>
</tr>
<tr>
<td>EcellReader</td>
<td>Ecell file I/O</td>
</tr>
<tr>
<td>Interaction</td>
<td>Protein Protein Interaction Analysis</td>
</tr>
<tr>
<td>KEGG</td>
<td>KEGG database access</td>
</tr>
<tr>
<td>Pathway</td>
<td>Pathway viewer</td>
</tr>
<tr>
<td>Serizawa</td>
<td>Ecell wrapper</td>
</tr>
<tr>
<td>G::Seq:</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>Align</td>
<td>alignment diffseq</td>
</tr>
<tr>
<td>AminoAcid</td>
<td>calc_pl amino_info peptide_mass</td>
</tr>
<tr>
<td>Codon</td>
<td>codon_compiler _distance_cu shannon_cu enc cbi icdi fop w_value cai phx bui aaui codon_usage</td>
</tr>
<tr>
<td>COMGA</td>
<td>COMGA_correlation COMGA_grapher</td>
</tr>
<tr>
<td>Consensus</td>
<td>base_information content, base_relative_entropy base_z_value</td>
</tr>
<tr>
<td>Eliminate</td>
<td>valid_cds eliminate_pat</td>
</tr>
<tr>
<td>FreeEnergy</td>
<td>RNAfold</td>
</tr>
<tr>
<td>GCskew</td>
<td>gcskew, cumgcskew, find_ori_ter, leading_strand, view_cds</td>
</tr>
<tr>
<td>ImaGene</td>
<td>ma_normalize ma_filter ma_rfilter</td>
</tr>
<tr>
<td>Markov</td>
<td>markov codon_markov</td>
</tr>
<tr>
<td>Operon</td>
<td>set_operon</td>
</tr>
<tr>
<td>ORF</td>
<td>longest_ORF find_identical_gene pseudo_atg</td>
</tr>
<tr>
<td>OverLapping</td>
<td>overlapping_finder</td>
</tr>
<tr>
<td>PatSearch</td>
<td>oligomer_counter find_seq palindrome find_dnaAbox</td>
</tr>
<tr>
<td>Primitive</td>
<td>complement translate</td>
</tr>
<tr>
<td>Tandem</td>
<td>find_tandem foreach_tandem_graphical LTR_search</td>
</tr>
<tr>
<td>Usage</td>
<td>rscu equitability cei</td>
</tr>
<tr>
<td>Util</td>
<td>genome_map molecular_weight plasmid_map gene_function_list</td>
</tr>
</tbody>
</table>

G-language
 Genome Analysis Environment
 Simulation Environment
Cross Platform GUI

- Perl + wxWindows and CGI

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

Commandline interactive interpreter for short analyses and trial.

- basic shell functions
- automatic logging
- persistent memory

perl compatible.
Research Cycle

Perl

G-language

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

generate perl script
pluggable dynamic
loader of
subroutines

generate perl script
http://www.g-language.org/

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

version 1

Bacteria Analysis System
K. Mori
K. Ikeda

C. DNA Analysis System
T. Matuzaki
Y. Kobayashi

Interpreter
D. Kyuma

Chi Sequence Analysis System
S. Nakamura

Comparative Genome Anal. System

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