G-language Project in 2009

○木戸信博，荒川和晴，大下和希，冨田勝
慶應義塾大学先端生命科学研究所

様々なオミクス情報が急速に蓄積される今日，バイオインフォマティクスはもはや分子生物学にとって不可分である．だが，数多くの解析ソフトウェアやデータベース、そして Bio* に代表される API が存在する一方，科学研究において中心的な役割を果たす試行錯誤のプロセスを支援するような統合解析環境の整備が望まれている．そこで，2001 年に慶應義塾大学先端生命科学研究所で立ち上げられた G-language プロジェクトでは特にこの試行錯誤のプロセスの効率化に重点を置き，汎用ゲノム解析ワークベンチ G-language Genome Analysis Environment をはじめとする様々なツールを構築してきた．本シンポジウムでは G-language プロジェクトの現状を，特に Perl ベースのライブラリ及びシェル，そして SOAP/REST の API を提供する G-language GAE と，多次元・多尺度なゲノム情報の閲覧を可能とする Genome Projector を中心に紹介する．これらのソフトウェアは http://www.g-language.org/ にてフリーで公開されている．

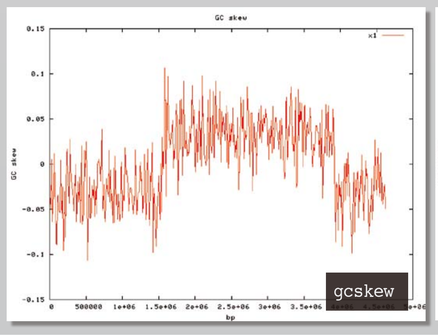
Contact
dokky@g-language.org

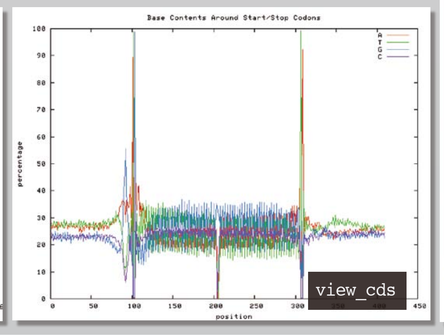
G-language GAE

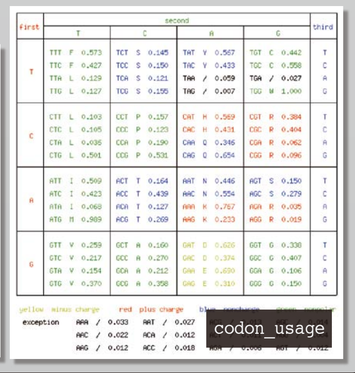
UNIX コマンドとの連携で，より快適に．

G-language Genome Analysis Environment (GAE) はバクテリアゲノムを解析するための 100 以上の関数をもつ統合解析環境である．G-language GAE ではキャッシングや構造の最適化により高速なデータアクセスを実現し，Perl ベースのシェルから操作することでインタラクティブかつプログラマブルな解析を可能にしている．また，Perl のモジュールとしても提供され，大規模な解析を行う際も強力にサポートを行うツールである．G-language GAE にはバクテリアゲノムの解析に特化した関数が多く実装され，例えば一本鎖 DNA 分子における G 含量と C 含量のバイアスを示す指標である GC skew をグラフ化する関数や，それを用いて複製開始地点・終結地点を予測する関数，その予測に基づいた複製進行方向の鎖（リーディング鎖）にコードされている遺伝子の一覧を取得する関数などがある．このように，G-language GAE では，バクテリアがもつ環状ゲノムとその複製の特徴を活かした解析を容易に行うことができる．また，G-language GAE の最新版でシェルに UNIX のパイプ機能が実装されたことより，G-language による解析の結果をパイプで grep に渡すなど，より快適に多彩なデータ処理を行えるようになっている．

```
G > $gb=load("ecoli")
Accession Number: NC_000913
  Length of Sequence :   4639675
           A Content :   1142228 (24.62%)
           T Content :   1140970 (24.59%)
           G Content :   1176923 (25.37%)
           C Content :   1179554 (25.42%)
              Others :         0 (0.00%)
          AT Content :     49.21%
          GC Content :     50.79%

G > gcskew($gb,-output=>"g")

G > find_ori_ter($gb)
find_ori_ter:
   Predicted Origin:   3923622
   Predicted Terminus: 1550412

G > @genes = genes_from_ori($gb)

G > foreach(@genes){ ... }

G > say togoWS('R00123') | head -n 3
ENTRY       R00123              Reaction
NAME        ADP aminohydrolase
DEFINITION  ADP + H2O <=> IDP + NH3
```







Web Services

REST/SOAP: アクセシビリティが向上．

BioHackathon 2009 において，我々のグループは G-language GAE の REST (Representational State Transfer) および SOAP インターフェイスを実装した．これらの実装により，G-language GAE の解析環境をインストール作業なしに利用することができるようになったほか，Perl 以外の各種プログラミング言語などからも利用できるようになり，他のツールとの連携も容易に行えるようになった．

■REST

G-language GAE では REST に準拠したインターフェイスを提供している．これは，解析関数とパラメータを URI で表現し，HTTP 経由でアクセスすることによって結果をテキストや画像で得られるというもので，ブラウザやその他 HTTP 通信をサポートする環境から解析を行うことができるインターフェイスである．http://rest.g-language.org は http://useG.jp でもアクセスが可能であり，解析を行うための URL は極力シンプルにデザインされている．

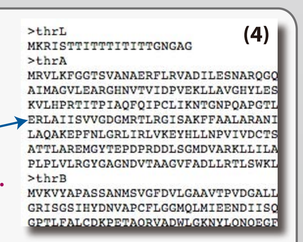
http://useG.jp/ecoli/recA/transtaion/ →

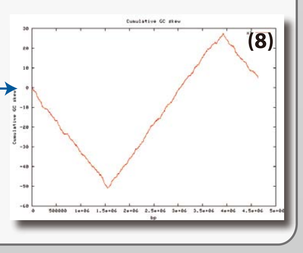
```
use G;
$gb = load("ecoli","no msg");
say $gb->{recA}->{translation};
```

□Bio::Glite

REST を利用した軽量版 G-language として，Perl のモジュール Bio::Glite を開発した．Bio::Glite は CPAN にて入手可能である．

1．フラットファイルからゲノムの情報を得る ... http://useG.jp/[species]/[gene]/[feature]/
(1) http://useG.jp/ecoli/（大腸菌のゲノム組成情報 :GenBank ファイルより）
(2) http://useG.jp/ecoli/recA/（*recA* 遺伝子のエントリーの詳細）
(3) http://useG.jp/ecoli/recA/start/ （*recA* 遺伝子の開始塩基位置）
(4) http://useG.jp/ecoli/*/translation/ （全遺伝子のアミノ酸配列 (FASTA 形式で出力 ））
2．ゲノムデータにアクセスする ... http://useG.jp/[species]/[gene]/[function]/([option1=value])/...
(5) http://useG.jp/method_list/gb (データアクセス関数一覧）
(6) http://useG.jp/NC_000913/*/before_startcodon/ （大腸菌全遺伝子の上流配列）
3．ゲノムの解析をする ... http://useG.jp/[species]/[function]/([option1=value])/...
(7) http://useG.jp/method_list/ (解析関数一覧）
(8) http://useG.jp/mgen/gcskew/cumulative=1/window=1000/（マイコプラズマ菌の累積 GC skew）
(9) http://useG.jp/mgen/gcskew/output=f/（マイコプラズマ菌の GC skew を CSV で出力）
4．他の関数を利用する ... http://useG.jp/[function]/([option1=value])/... # 関数一覧は (7) に同じ
(10) http://useG.jp/togoWS/C00001/ (統合 WS を利用して KEGG から C00001 のデータを取得）
(11) http://useG.jp/help/gcskew/（gcskew 関数のヘルプを表示）





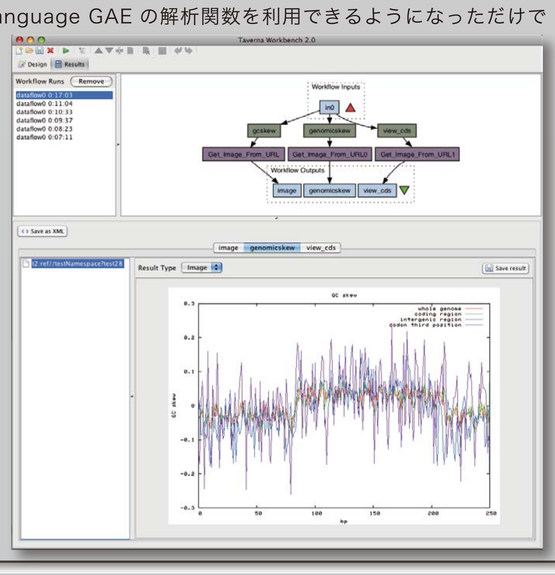
■SOAP

SOAP サービスとは HTTP を用いて XML 形式でメッセージを送受信するウェブインターフェイスで，WSDL (Web Service Description Language) にメソッドの詳細を記述し，そのフォーマットに沿ってメッセージを生成・解釈することで，異なるプラットフォーム間での通信を容易に実現するためのサービスである．我々は G-language GAE の解析関数に対応する WSDL ファイルと，SOAP による通信を処理する CGI を生成することで SOAP サービスを実装した．

これにより，SOAP をサポートする各プログラミング言語から G-language GAE の解析関数を利用できるようになっただけでなく，Taverna など SOAP に対応したツール連携用のソフトウェアからも利用することができ，他のゲノム解析環境と連携して G-language GAE を利用することができるようになった．

以下に，soap/wsdlDriver を利用した Ruby からの利用法と，Taverna を利用した解析例を示す．

```
#!/usr/bin/ruby
require 'soap/wsdlDriver'

wsdl = "http://soap.g-language.org/g-language.wsdl"
serv = SOAP::WSDLDriverFactory.new(wsdl).create_rpc_driver
serv.generate_explicit_type = true

print serv.gcskew("ecoli",'')
```



http://soap.g-language.org/g-language.wsdl

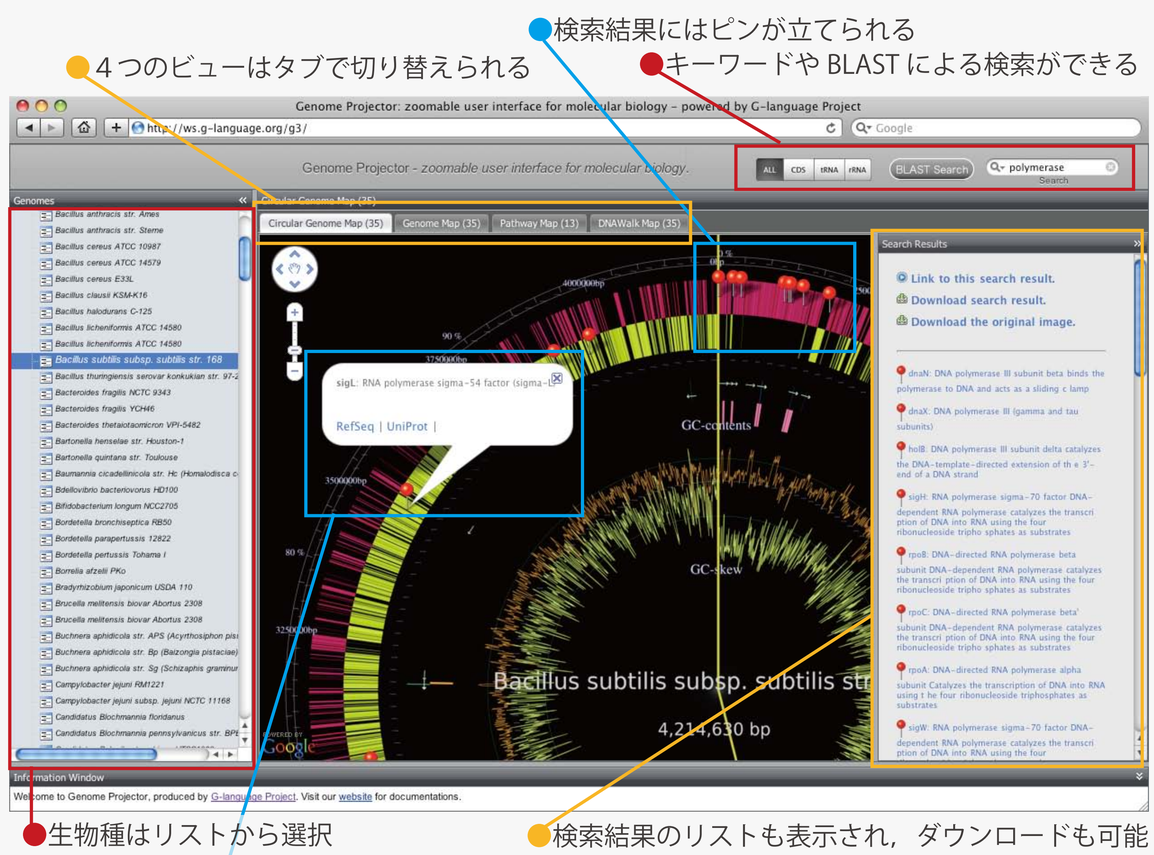
Genome Projector

ZUI を備えたシームレスなゲノムビューワー．

Google Maps は，ウェブアプリケーションでありながら，地図という巨大画像を高速にズームイン／アウトしながら閲覧し，かつ検索結果にピンを立て，ユーザが任意に情報をのせられるために注目を集めているソフトウェアである．ゲノムも同様に拡大／縮小しながら様々な視点から情報を眺めるべきものであるということに着眼し，本プロジェクトでは，Google Maps API を利用して ZUI (Zumable User Interface) を実装したシームレスなゲノムビューワー "Genome Projector" を開発した．

Genome Projector は４つのビューを備える．それは，ゲノムを環状に表現した Circular Genome Map view，従来のゲノムブラウザーと同様なレイアウトの Genome Map view, Roche Biochemical Pathway wall chart を基にした Pathway view，DNA の塩基配列をその組成を利用してベクターとして表現した DNA Walk view の４種類である．

Genome Projector では４つのビューとも ZUI を備え拡大／縮小をシームレスにできる上，AJAX を利用した非同期通信によりページ遷移せずにすべてのビューを横断して閲覧することができる．さらに，上部の検索ボックスにてキーワード検索をすることもでき，その結果に応じてマップ上にクリッカブルなピンが立てられる．そのため，検索結果の位置関係を把握しつつ外部データベースを参照することができるようになっている．また，あるビューでの検索クエリーは他のビューでも保持され，ビューを変更した際にも再検索せずに瞬時に結果が表示される．また，Genome Projector では API も提供しており，ユーザ独自のデータに基づいたビューワーを構築することができるようになっている．



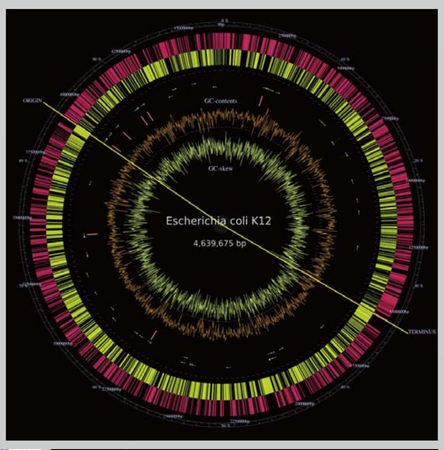
GenomeProjector の４つのビュー

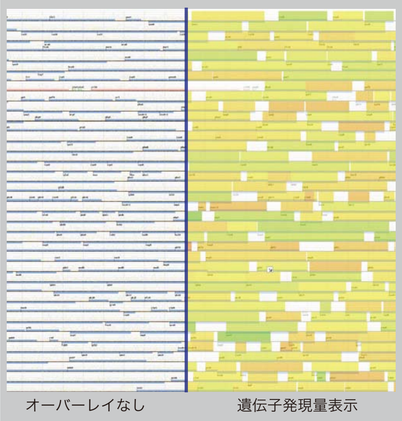
■Circular Genome Map view（右）
外側から順鎖 CDS 領域（赤），相補鎖 CDS 領域（黄），tRNA（緑矢印），rRNA（ピンク，オレンジ），GC content（茶），GC skew（黄緑）．予測された複製開始，終結地点も黄色の線で表現されており，この例ではそれぞれ 10 時，4 時の方向である．

■Genome Map view（下左）
線状にゲノムを表現したマップで，CDS，tRNA，rRNA の各領域がそれぞれ青，緑，赤のボックスで表現されている．CAI 値から予測された遺伝子発現量もオーバーレイとして表示できる（右半分）．

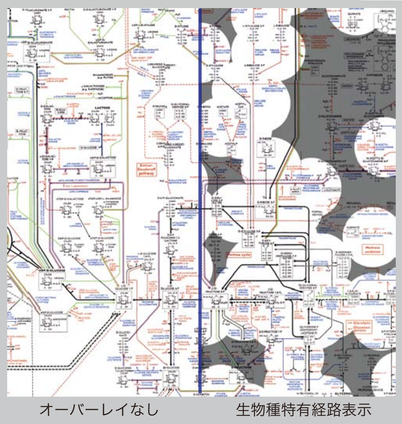
■Pathway view（下中）
Roche Biochemical Pathway wall chart を基にしたマップで，その生物種がもつ経路のみハイライトするオーバーレイ表示もある（右半分）．

■DNA Walk view（下右）
0,0 の座標を始点とし，塩基配列に応じて A で上，T で下，G で右，C で左に１ピクセルずつ進む．GC skew に強く傾向が見られる種では，複製開始地点から右に進み，終結地点から左に戻ってくるような図になる．

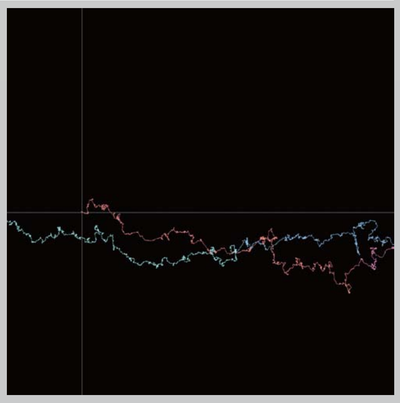




オーバーレイなし　遺伝子発現量表示



オーバーレイなし　生物種特有経路表示



http://ws.g-language.org/GenomeProjector/

References

Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y. and Tomita, M. (2003) G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics*, **19**, 305-306.
Arakawa, K. and Tomita, M. (2006) G-language System as a platform for large-scale analysis of high-throughput omics data. *Journal of Pesticide Science*, **31**, 282-288.
Arakawa, K., Suzuki, H. and Tomita, M. (2008) Computational Genome Analysis Using the G-language System. *Genes, Genomes and Genomics*, **2**, 1-13.
Arakawa, K., Tamaki, S., Kono, N., Kido, N., Ikegami, K., Ogawa, R. and Tomita, M. (2009) Genome Projector: zoomable genome map with multiple views. *BMC Bioinformatics*, **10**, 31.

Acknowledgement

G-language GAE の Web サービス開発は，DBCLS / OIST 主催の BioHackathon 2009 での成果物の一つである．
また，G-language Project は山形県及び鶴岡市の助成を受けて研究を行っている．

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