

# 比較ゲノム解析ツール「G-InforBIO」の開発



○田中 尚人<sup>1,2</sup>、阿部 貴志<sup>1</sup>、宮崎 智<sup>3</sup>、菅原 秀明<sup>1</sup>

- 1 情報システム研究機構 国立遺伝学研究所 生命情報・DDBJ研究センター
- 2 科学技術振興機構 バイオインフォマティクス推進事業
- 3 東京理科大学 薬学部

近年、細菌ゲノムを中心とした配列データおよびアノテーション情報が国際塩基配列データベース (INSD) を介して大量に公開されるようになってきた。ゲノム情報が蓄積されるにつれ、近縁種との比較解析がゲノム研究の主力となることは間違いないが、現在はそのための支援ツールの開発が進んでいない。そこで我々は、INSDに公開されているゲノム情報を利用した比較解析を可能としたツール「G-InforBIO」を開発した。このツールは以下の機能が備わっている。

- 1, ゲノム情報の管理・編集
- 2, Feature View: ゲノム上の遺伝子配列情報 (シンテニー) をグラフィカルに表示
- 3, Alignment View: Megablast による 2 ゲノム間の同源性検索およびその結果をグラフィカルに表示
- 4, VS Genome: ゲノムを subject とした核酸およびアミノ酸配列の blast 同源性検索
- 5, Blastclust: Blastclust (blast score-based single-linkage clustering) による核酸およびアミノ酸配列のクラスターング解析
- 6, ClustalW: 核酸およびアミノ酸配列の系統解析

G-InforBIO は WFCC-MIRCEN World Data Centre for Microorganisms のサイト (<http://www.wdcm.org/>) からダウンロードできる。

### ゲノム情報の管理、閲覧

**ゲノムデータの取り込み**

- Local DB: XML file を読み込む
- Import Flat File (FTP): 国際塩基配列データベースに登録されているゲノムデータを取り込み (DDBJ (塩基子 + seq) URL: <http://ftp.gene.dcc.ac.jp/>, NCBI (塩基子 + gdm URL): <http://www.ncbi.nlm.nih.gov/genome/MICROBES/complete.html>)
- 複数のデータを追加できる

**ゲノム情報のデータベース**

- 項目を選択
- ゲノムを選択
- データベース
- 項目の検索
- 項目の種類
- Feature: 遺伝子の種類 (ex. CDS, gene, rRNA, tRNA etc)
- Location: 領域指定 (ex. 1000-2000, 1000 番から 1999 番の間)
- Qualifier-key: 国際塩基配列データベースで統一されている識別子
- Qualifier-value: 種ごとに異なるデータ (Cf. <http://www.ncbi.nlm.nih.gov/projects/colibri/>)

**ゲノム情報の検索、閲覧**

**ゲノムマップ表示**

- アノテーション情報の詳細が表示される
- ゲノムマップ上に登録されたゲノム配列をマップ表示させるために Import from Search をクリックする
- 複数のゲノムがある場合を選択する

**配列データの切り出し**

- 指定した配列の切り出し
- コード領域を選択
- SEQ をクリック
- DNA およびアミノ酸配列がファイル形式で出力される
- ローテーションを指定入力でも同じとしたり可能

### 比較ゲノム解析

完全長ゲノム配列が公開されている *Cyanobacteria* 8 株の ORF セットより、共通 ORF を検出し、その ORF が系統分類学的指標となるか検証した。

ゲノム情報公開および解析中の *Cyanobacteria* 株の 16S rRNA 遺伝子配列に基づく系統関係

黒: 完全長ゲノムデータを INSD から公開

グレー: ゲノムプロジェクト進行中

Species: *Synechococcus* sp. PCC 7902, *Synechocystis* sp. PCC 6803, *Nostoc* sp. PCC 7120, *Prochlorococcus marinus* subsp. *marinus* CCMP1375, *Prochlorococcus marinus* subsp. *pastorikii* CCMP1986, *Prochlorococcus marinus* MIT 9313, *Synechococcus* sp. WH 8102, *Synechococcus* sp. PCC 7942, *Thermosynechococcus elongatus* BP-1, *Gloeobacter violaceus* PCC 7421.

### 1. Blastclust による ORF のクラスターング

データセット: 24,982 ORFs

430 clusters (複数の ORF により構成されたクラスター)

指標候補 ORF

クラスター構成配列の出力

系統解析

### 2. ClustalW による 共通 ORF の系統解析

指標候補 ORF: 4 セット

Cluster 14

Cluster 16

Cluster 1

Cluster 6

共通 ORF の系統解析により 16S rRNA 遺伝子による系統位置と異なる株があった。

### 3. Blast による ORF の同源性検索

(1) 近縁種の ORF と高い同源性を示す

(2) 系統的に離れた種の ORF と高い同源性を示す

重複や外來と思われる ORF を検証

ゲノムマップで周辺に transposase コード領域を確認

### 4. Megablast による 共通領域解析

3株のゲノム構造比較

共通領域の同源性比較

ヒット領域の詳細情報

ヒット領域

ゲノム構造の比較の結果、共通 ORF による系統関係を反映した。

*Cyanobacteria* 8 株の共通 ORF である thiamine biosynthesis protein 遺伝子および cytochrome b6 遺伝子が *Cyanobacteria* の系統分類の指標になりうる可能性がある。