微生物データ抽出・分類・同定システム InforBIO の開発

田中 尚人[1,2], 阿部 貴志[2], 宮崎 智[3], 菅原 秀明[2]

([1]JST BIRD, [2]遺伝研 生命情報・DDBJ, [3]東京理科大 薬)

（目的）
微生物の表現性状から遺伝子情報まで幅広い情報資源を論理的に統合してデータベース化、さらに解析するシステム InforBIO を我々は公開してきた。本システムには分類同定用解析プログラムを 7 種類搭載しているが、さらに、これらの結果を比較、連保して解析する新たな手法の開発を目的とした。

（方法および結果）
微生物株の様々な性状を利用したクラスタ解析手法として、コード化したデータを(数値)解析するデンドログラム解析と主成分分析、および配列データを(系統)解析する ClustalW がある。各手法における株の分布状況を比較し、上記の手法により構成されたクラスタの構成株の分布を最も反映する項目をコード化したデータ項目の中から検出する手法を開発した。この手法により、同定のための性状の絞り込みができる。また、ClustalW による系統解析で構成されたクラスタ構成メンバーの特異プライマーを自動的に検出する手法も開発した。本発表ではこれらの実例を挙げて紹介する。本システムは http://lilium.genes.nig.ac.jp/ からダウンロード可能である。

微生物情報のデータベース化

微生物情報をファイルで保管

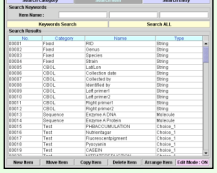
タブ区切りテキストファイル

XML ファイル

微生物情報の項目を二階層で管理

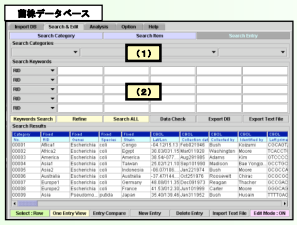
大項目: Category

小項目: Item



雲株の微生物情報データ管理

雲株データベース



キーワード検索（閲覧や解析のための抽出機能）

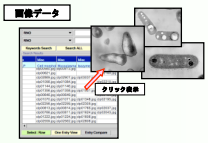
（1）Category キーワード
大項目に関する小項目を抽出

（2）Item キーワード
小項目に対して指定した性状を有する雲株を抽出

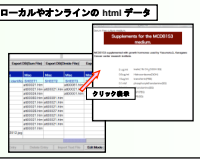
InforBIO で扱える微生物情報

表現性状データ（生理生化学的性状）や配列データに加えて以下のようなデータにも対応している

画像データ

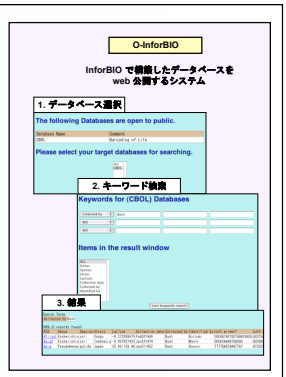


ローカルやオンラインの html データ



O-InforBIO

InforBIO で構築したデータベースを web 公開するシステム



微生物情報による分類同定解析

数値分類解析システム
小項目に対する性状をコード化して解析するプログラム
主に表現性状のデータを利用する

コード化するためのパラメータ設定画面

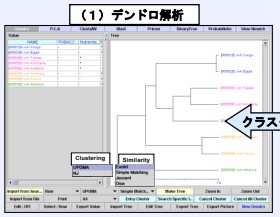
系統分類・配列解析システム
DNA, タンパク質配列を解析するプログラム

配列データのための項目設定画面

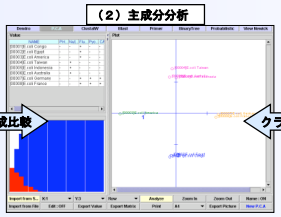
Category	Items	Type
Sequence	DNA or Protein name	Molecule

クラスタ解析

（1）デンドロ解析



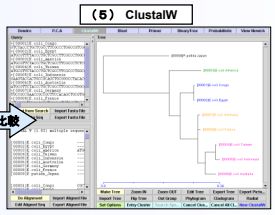
（2）主成分分析



クラスタ構成比較

クラスタ解析

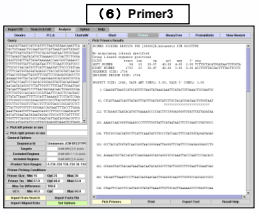
（5）ClustalW



クラスタ構成比較

プライマー解析

（6）Primer3



同定解析

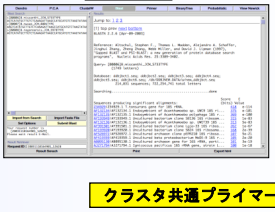
データベースを参考に同定用決定木（3）および同定スコア表（4）を作成

（3）決定木法

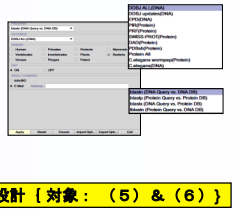
（4）確率的同定法

ホモロジー解析

（7）Blast



DDBJ サーバーで解析するための設定



クラスタ分別性状解析［ 対象：（1）（5）］

コード化したデータの分布をもとに、デンドロ解析またはClustalW で形成された複数のクラスタを分別する項目を検出する手法

クラスタ分別性状検索アルゴリズム

$$C = \{C_i : 1 \le i \le c_n\} : set\ of\ Clusters$$

$$V = \{v_j\} : set\ of\ values$$

$$p_i(v_j) : occurence\ probability\ of\ v_j\ in\ cluster\ C_i$$

$$Common\ Average = \frac{\sum_{C_i \in C} \sum_{v_j \in V} -p_i(v_j) \log p_i(v_j)}{c_n}$$

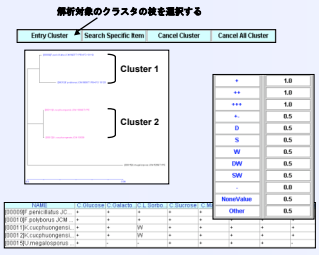
$$Differential\ Average = \frac{\sum_{C_i \in C} \sum_{v_j \in V} p_i(v_j) \log \frac{p_i(v_j)}{p_{i-1}(v_j)} + \sum_{C_i \in C} \sum_{v_j \in V} p_i(v_j) \log \frac{p_i(v_j)}{p_{i+1}(v_j)}}{2c_n}$$

Common Ave. （シャノンエントロピー）
クラスタ内の共通度

Differential Ave. （相関エントロピー）
クラスタ間のばらつき度

解析例

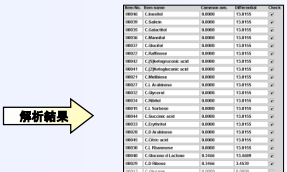
解析対象のクラスタの枝を選択する

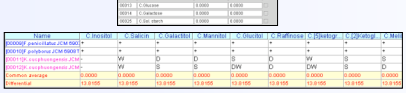


+	1.0
++	1.0
+++	1.0
+	0.5
D	0.5
S	0.5
W	0.5
DW	0.5
SW	0.5
	0.0
None/Value	0.5
Other	0.5

データをパラメータに置換し、その分布を解析する
上記の例では +, ++, +++ が同一としてカウントされる

解析結果





分別性状として有力な項目がリストに表示される

分別性状としての最適値
Common Ave.: 0
Differential Ave.: >0

クラスタ共通プライマー設計［ 対象：（5）＆（6）］

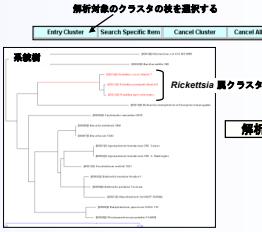
ClustalW で形成された１クラスタの共通プライマーを設計する手法

解析例

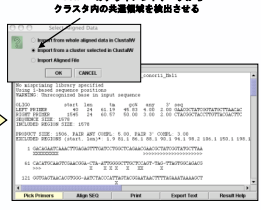
Rickettsia 属内共通の 16S rRNA 遺伝子のプライマー設計
および他の α-*Proteobacteria* との比較による特異性の確認

プライマー設計

解析対象のクラスタの枝を選択する

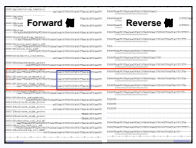


ClustalW のアライメントデータから
クラスタ内の共通領域を検出させる



解析結果

特異性確認



ClustalW のアライメントデータにForward, Reserse 個双方の配列および逆相補鎖があるかチェックし、表示される

左の例では、Forward 側のプライマーがデータセットの中では *Rickettsia* 属特異的なプライマーであることを示す