Plant GARDEN

種を超えた植物ゲノム情報統合のためのデータリンク基盤の構築

○市原寿子[1]、磯部祥子[2]、平川英樹[2]、原田大士朗[2]、 Ghelfi Andrea[2]、小原光代[2]、山田学[2]、白澤沙知子[2]、
中村保一[2]、田村卓郎[3]、杉原英志[3]、田畑哲之[2] 、中谷明弘[1]
（1. 大阪大学大学院医学系研究科, 2. かずさDNA研究所, 3. 筑波大学プレジション・メディスン開発センター）

概要

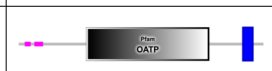
植物ゲノム統合ポータルサイトPlant GARDENにおいて、各植物ゲノムデータベースの配列データ間を繋ぐ基盤となるデータセットを構築している。配列データの連結は、遺伝子の**アミノ酸配列の類似性に基づく配列クラスタリング**により実施している。配列クラスタリングでは、同一遺伝子に由来する転写産物のバリアントが存在する場合、代表配列を選択して実施することが多い。Plant GARDENでは、公開データに可能な限り手を加えない方針を採用しており、代表ではなく全配列を対象とした配列クラスタリングを検証した。その結果、転写産物のバリアントの大部分は同一の配列クラスタに分類されたが、ごく僅かに異なるクラスタに分類されるものも観察された。また、Plant GARDENでは将来的な遺伝子配列の追加・更新作業を見据えた、類似配列データセットの構築方法を模索している。その方法として、植物の分類体系に沿って種、属、科などの各階層に属する生物種を一つの集合体として扱う方法や、配列プロファイルの利用が挙げられる。

転写産物のバリアント：代表配列を選択せずに全配列でクラスタリングを実施

- 明記のない植物ゲノムデータが混在する。
- 大部分のバリアントは同じ配列クラスタに分類された。
- バリアント同士が異なる複数のクラスタに分類される遺伝子が僅かに観察された。

植物名	遺伝子数	配列数	バリアント無	バリアント有	分類先が複数
A. thaliana (シロイヌナズナ)	27,655	48,359	16,960	10,695	206 (0.7%)
R.sativus (ダイコン)	46,373	46,373			
G.max (ダイズ)	55,897	88,412	40,629	15,268	560 (1.0%)
A. hypogaea (ラッカセイ)	66,872	84,714	55,024	11,848	330 (0.5%)
L.japonicus(2.5) (ミヤコグサ)	21,808	21,808			
L.japonicus(3.0) (ミヤコグサ)	39,734	48,105	33,362	6,372	66 (0.2%)
S.lycopersicum (トマト)	35,768	35,768			
V.vinifera (ヨーロッパブドウ)	29,971	29,971			

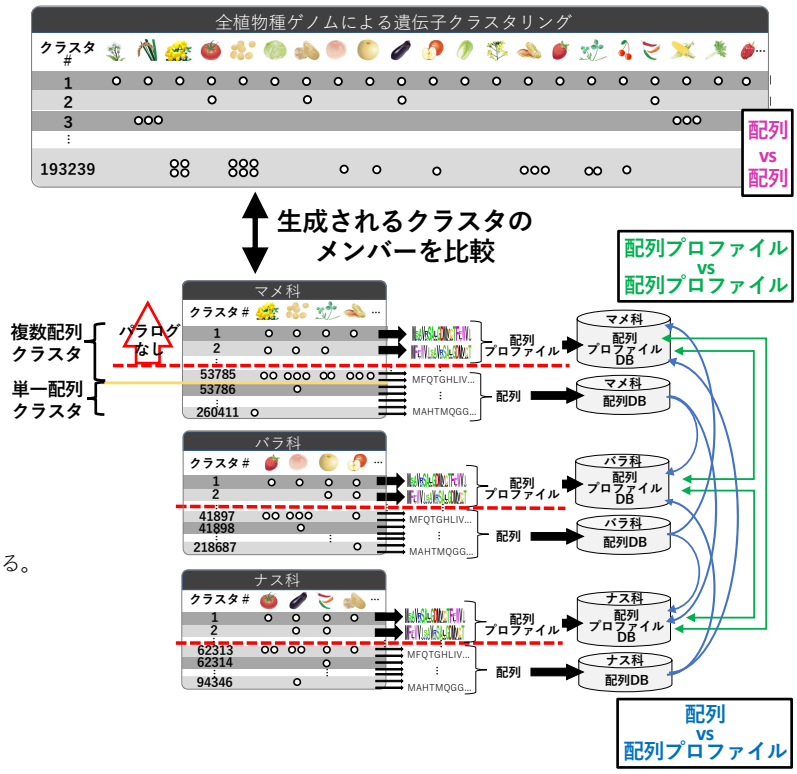
異なるクラスタに分類されたバリアントの例: 遺伝子IDAT2G22730（シロイヌナズナ）
遺伝子クラスタリングでは３つの配列クラスタに分類された。
ドメイン構造に基づく機能予測では異なる機能アノテーションが検出された。

クラスタID	メンバー配列の数	メンバー配列の遺伝子ID	DB検索によるドメイン構造予測	機能アノテーション
197856	4	AT2G22730.1, AT2G22730.2, W8PLG7.1(ラッカセイ), JM8NV3.1(ラッカセイ)		Probable sphingolipid transporter spinster homolog 3
195038	4	AT2G22730.3, AT2G22730.6, 6HDK8W.1(ラッカセイ), X38HVK.1(ラッカセイ)		Major facilitator superfamily protein
2862	2	AT2G22730.7, VIT_00s0282g00040.t01 (ヨーロッパブドウ)		記述なし

類似配列データセットの構築方法（逐次追加方法）の検討

操作手順

1. アミノ酸配列の相同性検索と類似度に基づいた遺伝子クラスタリング
   （ProteinOrtho, https://www.bioinf.uni-leipzig.de/Software/proteinortho/ ）
   a. 対象生物種の全配列間
   b. 共通植物科（イネ科、マメ科、バラ科等）内の配列間
2. 1-bの結果に基づいたマルチプルアラインメントの生成
   (Mafft, https://mafft.cbrc.jp/alignment/software/)
3. 2からアミノ酸配列プロファイルの生成
   (HAMMER, http://hmmer.org/)
4. 配列プロファイルを用いたデータベース検索
   a. 配列プロファイル vs 配列 (HMMER)
   b. 配列プロファイル vs 配列プロファイル (PRC, Profile Comparer , http://www.ibi.vu.nl/programs/prcwww/info.php)
5. 1-a、4-a、 4-bで得られたクラスタの比較

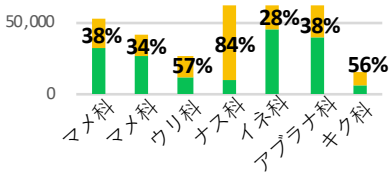


配列プロファイル vs 配列

検索にかかる時間が1/6 ~ 1/8に短縮
同じクラスタにアサインされた一致率は80~84%

クラスタが一致しなかったデータについて

種間・種内での類似度が低い遺伝子群ほど一致度が下がる傾向を示した。
複数配列から構成される全クラスタでプロファイルを生成しているので、
今後、プロファイル生成に採用するクラスタに閾値を設定して検証を進める。



パラログを含まない複数配列クラスタの割合
(プロファイルを生成するのに使用するクラスタ候補)

本プロジェクトは、科学技術振興機構（JST）・バイオサイエンスデータベースセンター（NBDC）の統合化推進プログラムの助成により実施しました。