

ゲノム配列から同定した遺伝子の機能予測を最速で行うプログラムを 開発しました

～Hayai-Annotation Plants～

5月15日に *Bioinformatics* 誌にてオンライン公開

令和元年5月27日

公益財団法人 かずさDNA研究所

- ◇ かずさDNA研究所は、植物のゲノム解析において重要な、遺伝子配列から遺伝子の働きを推定するプログラム「Hayai-Annotation Plants」を開発しました。
- ◇ ゲノムを解析し他の植物種と比較するには、配列データを解析してどこにどのような遺伝子があるかを注釈づける（アノテーション）作業が必要です。
- ◇ これまでアノテーション作業には大型の計算機が必要でしたが、開発したプログラムは通常のパソコンレベルのスペックで解析できます（現在は Mac 版 OS にのみ対応）
- ◇ 計算は従来法に比べて非常に高速で、さらに精度の高い結果を得ることができます。
- ◇ 研究成果は、国際科学雑誌 *Bioinformatics* に5月15日付でオンライン公開されました。

<報道に関すること>

公益財団法人かずさDNA研究所 広報・研究推進グループ

TEL : 0438-52-3930

<研究に関すること>

公益財団法人かずさDNA研究所 植物ゲノム・遺伝学研究室

室長 磯部 祥子 (いそべ さちこ) TEL : 0438-52-3935

1. 背景

次世代シーケンサーなどの DNA 配列解析技術の向上により、多様な植物種のゲノム配列が解読されています。かずさDNA研究所でも近年、トマトやイチゴなど多くの実用作物のゲノムの解析を行っています。実用植物の種・品種間のゲノム配列の違いによって多くの形質の差異が生じることから、それらのデータを比較して育種に利用するには、配列データを解析してどこにどのような遺伝子があるかを注釈づける(アノテーション)作業が必要です。

アノテーションには、構造アノテーションと機能アノテーションがあります。構造アノテーションとはゲノム配列から遺伝子として機能する配列の部分を見つけ出すことで、機能アノテーションとは見つけ出された遺伝子配列がどのような機能をもっているのかを推定することです。機能アノテーションでは、これまでに世界中で解析された様々な生物種の遺伝子のもつ情報が格納されているデータベースを用いて、配列の類似性などから調べたい遺伝子がどのような機能を持っているかを推定します。検索の対象となる配列数が膨大であることから、これまでは大型計算機サーバーを用いて計算を行う必要がありました。

今回開発した「Hayai-Annotation Plants」は、機能アノテーションをつけるためのプログラムで、植物を対象としています。各遺伝子に対して、1) タンパク質の名前、2) 遺伝子オントロジー (GO ; 遺伝子の機能や細胞内局在を示すタグのようなもの) : 3 つの GO カテゴリーから推定されたもの、3) EC 番号 (酵素の機能を分類するためのもの)、4) 機能推定の精度レベル、5) 推定の根拠となる情報の由来、を速く、かつ高精度で付与することができます。これらのデータは主にゲノム解読の際に得られた遺伝子配列の機能を推定する際に利用されます。

かずさDNA研究所では、これらの情報も含めたさまざまな植物のゲノム情報を集めたポータルサイト、Plant GARDEN (Genome And Resource Database Entry: https://plantgarden.jp/app/pc_01.html) を運営し、統合データベース PGDB (Plant Genome Database Japan: <http://pgdbj.jp/index.html>) において検索に有効なリソースの整備を行っています。

2. 研究成果の概要と意義

- ① 高速な計算を実現するために、植物を対象としたアノテーションに特化させました。
- ② データベースの配列セットを分割し、検索を並列処理することによってラップトップ型パソコンでの処理を可能にしました。
- ③ 機能アノテーションは次の4段階で実施し、推定された機能の精度を評価することができます。1) 対象とする種でアミノ酸配列が存在している、2) 転写産物の配列が存在している、3) モデル植物で類似の遺伝子配列が存在している、4) アミノ酸配列の推定のみ、の順で評価し、これにより精度の高い結果を得ることができました。
- ④ シロイヌナズナの遺伝子配列データセット (約3万遺伝子) を用いたテストでは、およそ5分で解析が終了しました。

3. 将来の波及効果

- ① 大型計算機サーバーや特別な専門性がなくても遺伝子アノテーションを実施できることから、大学などのアカデミアのみならず、企業の研究開発部署においても多様な植物種の遺伝子機能推定が進むことが期待されます。
- ② 植物のゲノム解析が加速され、多様な植物種でゲノム情報をもとにした育種が可能になります。

この研究は、かずさ DNA 研究所、および、JST ライフサイエンスデータベース統合推進事業の助成によって行われました。

論文タイトル： Hayai-Annotation Plants: an ultra-fast and comprehensive functional gene annotation system in plants.

著者： Andrea Ghelfi, Kenta Shirasawa, Hideki Hirakawa and Sachiko Isobe.

掲載誌： *Bioinformatics*

DOI： 10.1093/bioinformatics/btz380

参考資料

The screenshot displays the Hayai-Annotation Plants web application interface. The top navigation bar includes the logo and the name 'Hayai-Annotation Plants' by Kazusa DNA Research Institute. The main content area is divided into several sections:

- Search Filters:**
 - Type of Alignment:** Local (selected), Global.
 - Type of Algorithm:** Protein Existence Level (selected), Alignment Score.
 - Max hits per query:** 1
 - Minimum Sequence Identity (%):** 80
 - Evaluate 1e-:** 6
 - Minimum Query Coverage (%):** 80
- Upload FASTA File:** A 'Browse...' button with the text 'No file selected'.
- Submit:** A button to execute the search.
- Download:** A button to download the results.

Below the filters, there is a table showing search results. The table has columns for query, seqID, length, msa, gaps, startquery, endquery, starttarget, endtarget, evalue, score, gene_name, Protein_Evidence, EC, evidence_type, GO_BP, GO_BP_name, GO_MF, GO_MF_name, GO_CC, and GO_CC_name. The first few rows of the table are as follows:

query	seqID	length	msa	gaps	startquery	endquery	starttarget	endtarget	evalue	score	gene_name	Protein_Evidence	EC	evidence_type	GO_BP	GO_BP_name	GO_MF	GO_MF_name	GO_CC	GO_CC_name
4 AT1G01562.2 DCL1_ARATH	99.8	1910	2	1	1	1910	1	1909	0	3828.2	Endonuclease Dicer domain 1	PE 1	3.1.26.-	ISA	GO:000366	RNA processing	GO:000365	endonuclease II activity	GO:0005737	cytoplasm
5 AT1G01303.1 IPYR1_ARATH	100	212	0	0	1	212	1	212	1.1e-119	431.8	Soluble inorganic pyrophosphatase 1	PE 1	3.6.1.1	IDA	GO:000676	phosphate-containing compound metabolic process	GO:000589	magnesium ion binding	GO:0005839	cytosol
6 AT1G01380.1 LHY_ARATH	100	845	0	0	1	845	1	845	0	1298.5	Protein LHY	PE 1		IDA	GO:0042754	negative regulation of circadian rhythm	GO:0003703	DNA-binding transcription factor activity	GO:0005634	nucleus
12 AT1G01120.1 KCS1_ARATH	100	529	0	0	1	529	1	529	0	1052	3-hydroxy-CoA synthase	PE 1	2.3.1.199	IDA	GO:000689	fatty acid biosynthetic process	GO:0102768	very-long-chain-3-ketolase-CoA synthase activity	GO:0005768	endoplasmic reticulum
14 AT1G01145.3 CPN9_ARATH	99.3	453	1	1	1	451	1	453	2.2e-258	892.8	serine/threonine-protein kinase 3	PE 1	2.7.11.1	IDA	GO:0005070	potassium ion homeostasis	GO:0005504	ATP binding	GO:0005737	cytoplasm
22 AT1G01220.1 PKSP_ARATH	100	1056	0	0	1	1056	1	1056	0	2115.5	8Functional nucleoside kinase synthetase	PE 1		IDA	GO:0042392	GDP-L-Adenosine salvage	GO:0047240	nicotinamide guanylate transferase activity		
27 AT1G01280.1 BNC13_ARATH	100	890	0	0	1	890	1	890	0	1177.2	Transcription factor bHLH13	PE 1		IDA	GO:000369	transcription, DNA-templated	GO:0044212	transcription regulator region DNA binding	GO:0005634	nucleus
28 AT1G01285.1 CTD42_ARATH	100	510	0	0	1	510	1	510	0	1044.3	Cytoskeleton P450 703A2	PE 1	1.14.14.130	IBA	GO:0010584	protein homeostasis	GO:0002702	fatty acid in-chain hydrolyase activity	GO:0018020	membrane