Rorippa aquatica 遺伝子情報データベースの構築

平成 28 年 4 月 19 日受付

坂 本 智 昭*
木 村 成 介*

要 旨

Rorippa aquatica は環境に応答した葉形の変化やホルモンフリー条件下での葉断片からの植物体再生といった興味深い形質を示す植物である。しかしながら，本種のゲノムおよび遺伝子情報は整備されておらず，現在は次世代シーケンサーを用いた解析によってそれらの情報の取得が進行中である。現段階でも得られた遺伝子情報は膨大であり，その解析結果も多岐に渡っている。そこで本研究では取得した遺伝子情報およびその解析結果を遺伝子情報データベースとして統合するとともに，収集したデータを簡易に検索・閲覧することを可能にするインターフェイスの整備を試みた。

キーワード：*Rorippa aquatica*，de novo アセンブル，データベース，SQLite，Perl/CGI

1. はじめに

Rorippa aquatica は北米原産のアブラナ科の半水生の植物である。この植物を水中で育成すると気中で育成した場合と異なる形の葉を形成する。また，これまでの研究により気中においても温度や光強度といった周囲の環境に応じて葉の形態が変化することが明らかになっている。特に温度の影響は大きく，30℃の高温条件では楕円形の単葉を形成するが，温度が低くなるに従って葉の切れ込みが大きくなり，小葉も細くなっていく。このような環境に応答した葉の形態の変化を示す *R. aquatica* は葉の形態形成, 環境応答およびそれらをつなぐシグナル伝達系の解析において有用な実験材料である。

R. aquatica は外部から植物ホルモンを添加しなくても葉の断片から植物体を再生させるという特徴を持ち合わせている。これまでの研究から葉断片からの植物体再生にはオーキシンが関与していることが示唆されている。オーキシンの合成，輸送といった植物ホルモンの作用経路と植物体の分化・再生に至る経路の関連を解析するのに重要な形質であると考えられる。

また，本研究室では *R. aquatica* の 2 つの地域系統が維持されている。この系統間では環境に対する葉形の応答性に大きな違いが見られる。一方の系統では環境に応じて葉形を大きく変化させるが，もう一方の系統では環境を変化させた場合もほとんど葉形が変化しない。また，葉形以外にも様々な形質において系統間で差異が見られ, これらは系統間の遺伝的差異によって生じることが示唆される。

しかしながら，本種はモデル生物ではないためゲノムおよび遺伝子情報が整備されていない。その

* 京都産業大学総合生命科学部

ため，次世代シーケンサーを利用してゲノムおよび遺伝子情報の取得が進められている。また，様々な環境への応答を網羅的に解析するため，RNA-seq によるトランスクリプトーム発現解析も並行して行われている。

これまでの解析として茎頂部分の de novo RNA-seq 解析が行われている。気中条件下で温度と光強度を変化させて育成した植物体から葉原基を含む茎頂を取り出して全 RNA を抽出し RNA-seq ライブラリを作成した。作成したライブラリを用いて次世代シーケンサーにより mRNA の断片配列情報を取得した。そして得られた mRNA 断片配列情報をもとに de novo アセンブルを行い，サンプル中に含まれる全転写産物配列（トランスクリプトーム）情報を得た。さらに得られた転写産物情報についての解析を行ない各転写産物のアノテーションを行った。

このように次世代シーケンサーデータを用いた解析によって様々なデータが取得されているが，そのデータ量は膨大であり出力された解析結果をそのままの状態で扱うには困難が伴う。また，並行して行われる解析は相互に関連しており，より詳細な解析のためにはそれぞれの解析を関連付けて閲覧できる環境が必要である。そこで本研究では *R. aquatica* を用いた解析結果を網羅的に収容するデータベースを構築するとともに，それらの結果を統合して検索および閲覧を行うためのインターフェイスの整備を行った。

2. *R.aquatica* データベースの構築

データベース言語には管理・移植の簡易さから SQLite を選択した。単一のデータベースファイルを作成し，それぞれの解析結果ごとに table（テーブル）を設定し，必要なデータのインポートを行った。

2.1 de novo アセンブル転写産物配列データ

de novo アセンブルにより 132556 個の転写産物配列（コンティグ）が得られている。de novo アセンブルでは系統間での比較を可能にするため，2 つの地域系統から得られた次世代シーケンサーデータを混合して解析に使用した。以降の解析はこの配列情報を基準にして行われている。得られた転写産物配列情報は multi-fasta 形式で出力されており，このデータから必要な情報を抽出して転写配列テーブルとしてデータベースに取り込んだ。

表 1　転写配列テーブルのフォーマット

カラム番号	カラム名	カラム内容詳細
1	Seqname	コンティグ配列名。本テーブル内では同一名のデータは含まれない。(例：c10083_g1_i1)
2	Sequence	コンティグ全長塩基配列

2.2 転写産物アノテーションデータ

アセンブルで得られた転写配列は塩基配列情報のみを持っており, その機能については未知である。そのため既知の遺伝子データベースを利用してアノテーション（注釈付け）を行った。

2.2.1 blastX による相同遺伝子の探索

既知のアミノ酸配列データベース Non-redundant protein sequences from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq (nr) を用いて blastX による mRNA 配列 - アミノ酸相同性検索を行い, コンティグと相同な遺伝子を探索した。blastX 結果は xml 形式で出力され，相同性のあるアミノ酸配列名，マルチプルアライメント結果等の情報を含む。データベースには表のように相同性のあるアミノ酸名，および e-value を取り込んだ。このテーブルでは 1 つのコンティグにつき複数の blast 検索結果を含んでいる。

表 2　blastX 検索結果テーブルのフォーマット

カラム番号	カラム名	カラム内容詳細
1	Seqname	コンティグ配列名。転写配列テーブルに含まれるものと相同。本テーブル内で同名データを含む場合あり。
2	Blast_hits	コンティグ配列と相同性を示した既知のアミノ酸配列名。nr データベースでは配列名にアミノ酸 ID，アミノ酸名，生物種名等を含む。
3	Evalue	Balstx で出力される e-value。小数または指数形式で示される。

2.2.2 Gene ontology（GO）の関連付け

blastX 検索によって見つかった相同な遺伝子が機能的に既知である場合，その遺伝子には biological process，cellular component，molecular function のいずれかの情報を示す GO term およびその ID である GOID が割り振られている場合がある。blastX によって見つかった既知の相同遺伝子の GO 情報を基にコンティグの Gene ontology の関連付けが行われた。GO 情報はコンティグ配列名とそれに関連付けられ GO ID からなり，それらの情報をデータベースに GO 情報テーブルとして取り込んだ。GO の関連付けは単一の配列に複数の GO ID が割り振られることがあるため, GO 情報テーブルには 1 データにつき 1 つのコンティグ配列名と 1 つの GO ID をとりこみ，複数の GO ID が関連付けられた場合別々のデータとなるように成形して取り込みを行った。

表 3　GO 情報テーブルのフォーマット

カラム番号	カラム名	カラム内容詳細
1	Seqname	コンティグ配列名。転写配列テーブルに含まれるものと相同。
2	Go	Gene ontology の関連付けによりコンティグに割り当てられた GO ID（例：GO:0000001）

上記の GO 情報テーブルは GO ID のみを含んでいるが，Gene Ontology では 1 つの GO ID およびそれと対応する GO term，それらが属する namespace が定義されている。それらの定義については Gene Ontology Consortium（http://geneontology.org/）によって定められている [1]。公開されているそれらの定義情報も同様にテーブルへの取り込みを行った。2 つのテーブルの情報は必要に応じて結合され出力される。

表 4　GO 定義テーブルのフォーマット

カラム番号	カラム名	カラム内容詳細
1	Go	GO ID。上記 GO 情報テーブルにふくまれるものと相同。
2	Namespace	各 GO が属する namespace。biological process，cellular component，molecular function のいずれかに分類されている。
3	Name	GO term

3. データベースアクセスインターフェイスの整備

以上の通りデータベースを構築することが出来た。しかし，データベース単体では，検索・閲覧を行うためには SQLite 言語への理解が必要とされる。また，複数環境でのデータベース閲覧にはデータベースの複製および移設が必要になる。そこで, データベースをサーバ上に置くことで, ネットワーク上の複数クライアントからのアクセスを可能にするとともに，インターネットブラウザからアクセス可能な検索インターフェイスを構築することで簡易なアクセスを可能にした。検索インターフェイスは検索フォームを含む HTML により構成された検索ページと Perl/CGI によって動的に作成される検索結果ページからなっている。

検索ページには blastX 検索結果のキーワード検索，コンティグ名のリストによる blastX 結果の抽出，GO 関連付け結果の検索の 3 つの検索フォームを設置した。

3.1 blastX 検索結果キーワード検索

blastX 検索結果のキーワード検索は相同アミノ酸配列名のキーワードによる検索を行うことができる。複数キーワードでの検索は全てのキーワードを含む AND 検索を行う。本検索ではキーワードによる検索の他に検索結果のフィルタリングのためのオプションを用意した。1 つは e-value によるフィルタである。blast 検索では相同性の確度として e-value が出力され，この値が小さいほど確度は大きい。このフィルタによってより確度の高い検索結果のみを出力することが可能になった。2 つ目は出力数のフィルタである。データベースに登録されている blastX 結果には 1 つのコンティグにつき e-value が小さい順に 20 個までの検索結果が登録されているが，閲覧性の向上のため上位のデータのみの出力に対応した。この検索により既知の機能遺伝子と相同な *R. aquatica* のコンティグ配列の取得を可能にした。

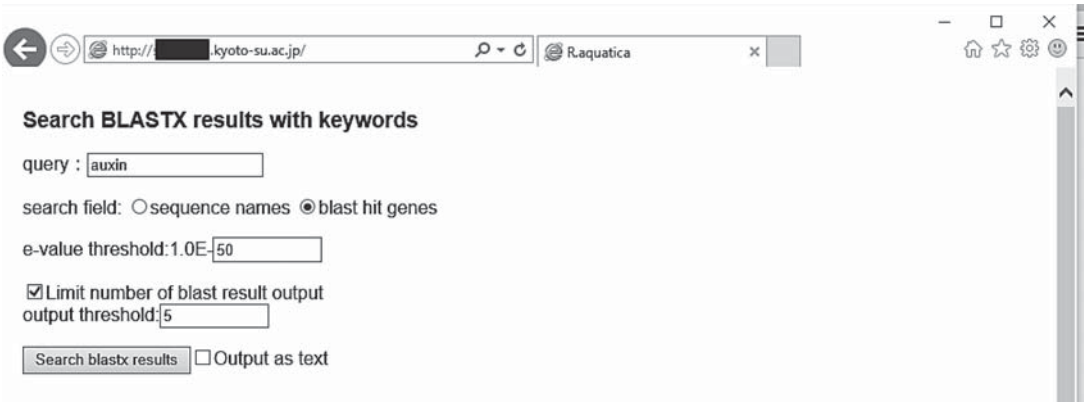


図 1　blastX 検索結果キーワード検索画面

3.2 コンティグ名のリストによる blastX 結果の抽出

発現量解析などの下流の解析およびフィルタリングによってコンティグのリストが出力される場合がある。その場合，より詳細な解析や特に注目すべきコンティグのフィルタリングには各コンティグの機能情報が必要になることが考えられる。本検索フォームでは入力されたそれぞれのコンティグに対して，相同性のある遺伝子のリストを返す。本検索でも上記の検索と同様に e-value と出力遺伝子数によってフィルタリングすることが出来る。

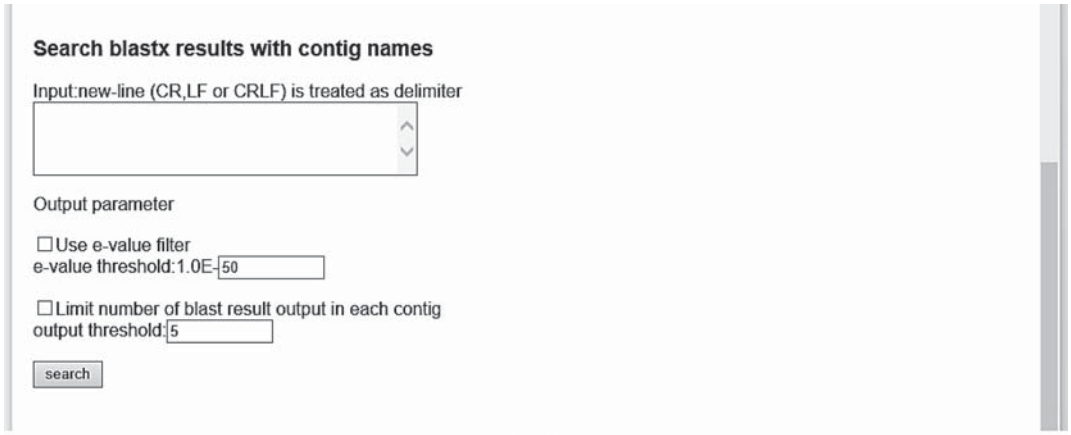


図 2　コンティグ名のリストによる blastx 結果抽出画面

3.3 GO 関連付け検索

発現解析や GO enrichment 解析によって特定の機能や作用を持つコンティグ群が抽出される場合がある。特定の機能によるコンティグの検索は上記のキーワード検索では不十分であると思われる。そのため, GO による検索を実装した。これにより GO 関連付けにより機能が推測されているコンティグを網羅的に抽出することができる。

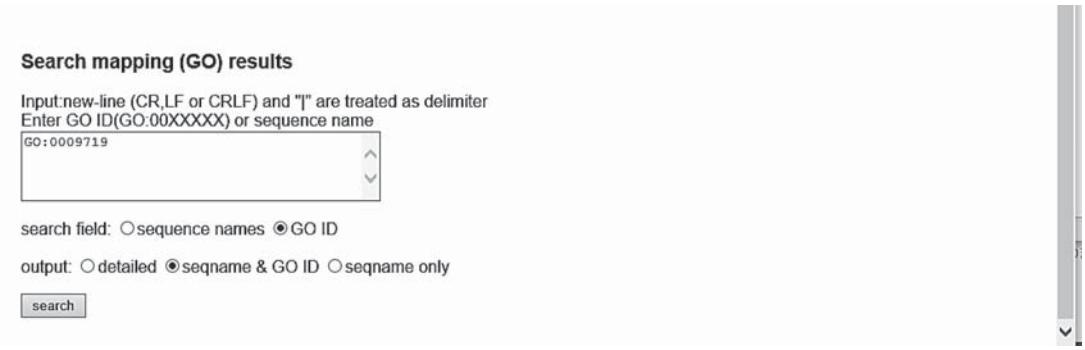
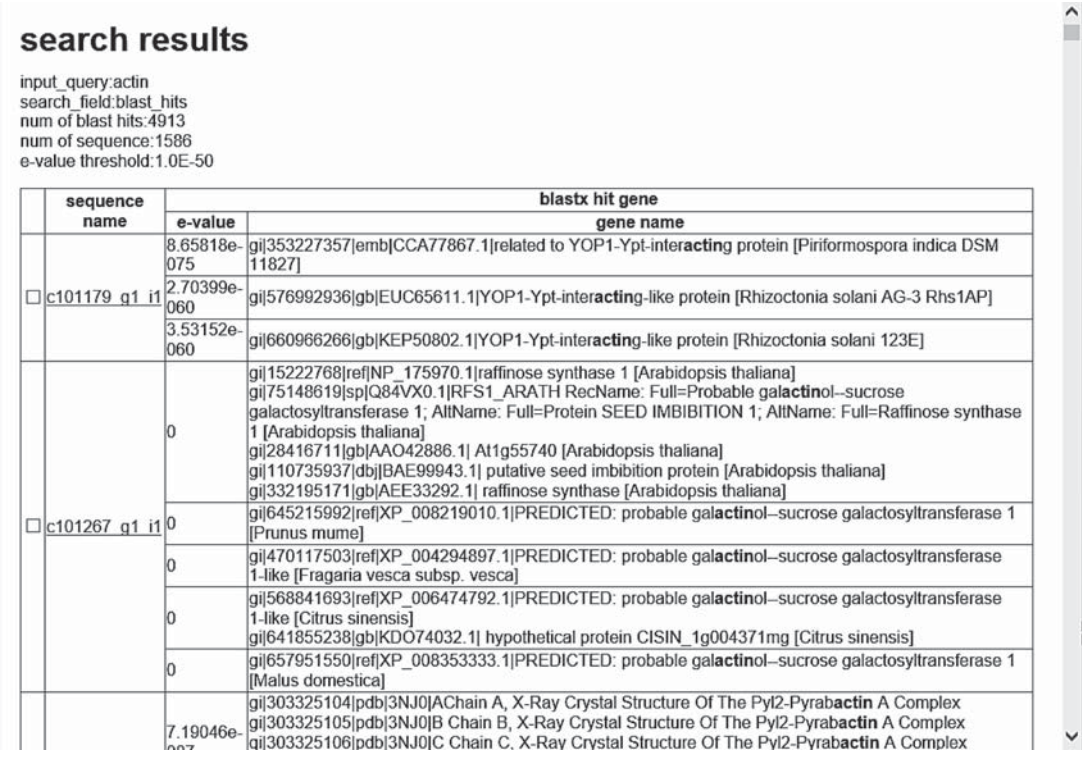


図 3　GO 関連付け検索画面

3.4 検索結果ページ

各検索フォームから出力される検索結果ページは主にコンティグ名，検索結果およびコンティグ名毎のチェックボックスで構成されている。

コンティグ名には後述するコンティグ詳細ページへのリンクが設定されている。また，チェックボックスへチェックを入れることにより，チェックしたコンティグの塩基配列を fasta 形式で出力できる。



search results

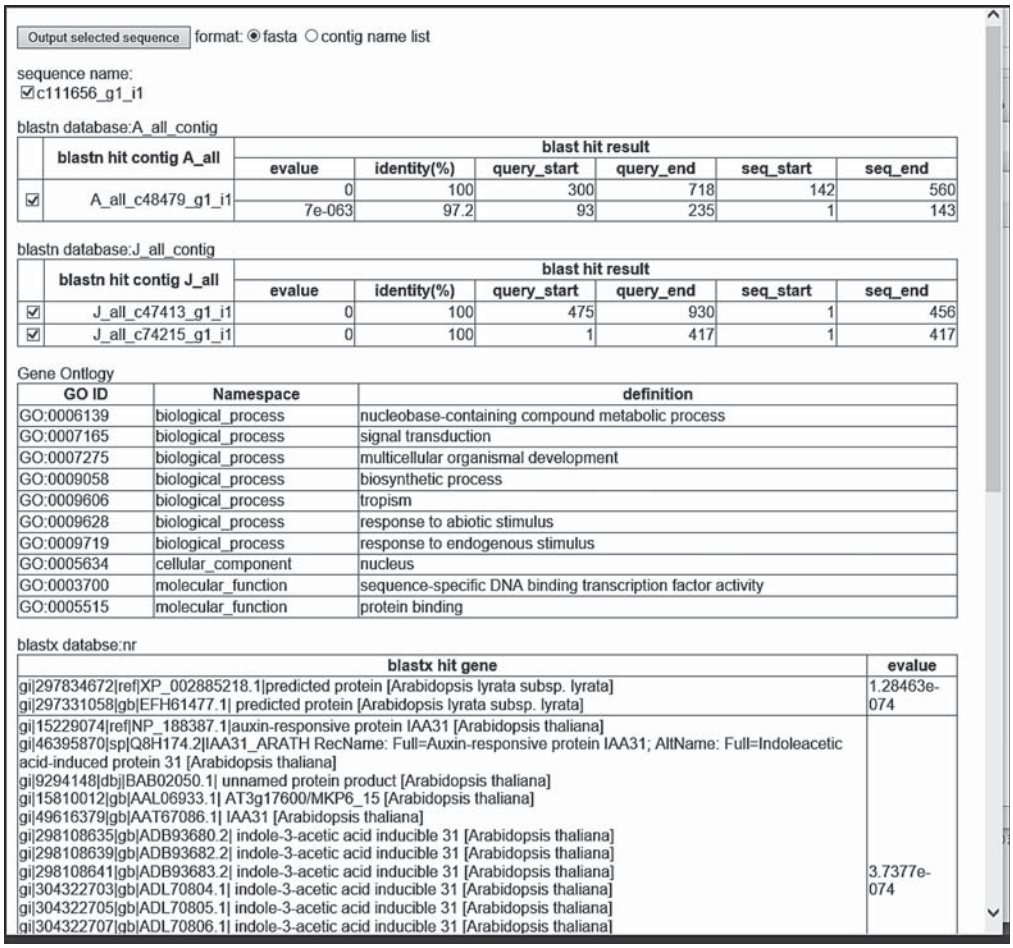
input_query:actin
search_field:blast_hits
num of blast hits:4913
num of sequence:1586
e-value threshold:1.0E-50

	sequence name	blastx hit gene	
		e-value	gene name
☐	c101179_g1_i1	8.65818e-075	gi|353227357|emb|CCA77867.1|related to YOP1-Ypt-inter**acting** protein [Piriformospora indica DSM 11827]
		2.70399e-060	gi|576992936|gb|EUC65611.1|YOP1-Ypt-inter**acting**-like protein [Rhizoctonia solani AG-3 Rhs1AP]
		3.53152e-060	gi|660966266|gb|KEP50802.1|YOP1-Ypt-inter**acting**-like protein [Rhizoctonia solani 123E]
☐	c101267_g1_i1	0	gi|15222768|ref|NP_175970.1|raffinose synthase 1 [Arabidopsis thaliana] gi|75148619|sp|Q84VX0.1|RFS1_ARATH RecName: Full=Probable gal**actin**ol--sucrose galactosyltransferase 1; AltName: Full=Protein SEED IMBIBITION 1; AltName: Full=Raffinose synthase 1 [Arabidopsis thaliana] gi|28416711|gb|AAO42886.1| At1g55740 [Arabidopsis thaliana] gi|110735937|dbj|BAE99943.1| putative seed imbibition protein [Arabidopsis thaliana] gi|332195171|gb|AEE33292.1| raffinose synthase [Arabidopsis thaliana]
		0	gi|645215992|ref|XP_008219010.1|PREDICTED: probable gal**actin**ol--sucrose galactosyltransferase 1 [Prunus mume]
		0	gi|470117503|ref|XP_004294897.1|PREDICTED: probable gal**actin**ol--sucrose galactosyltransferase 1-like [Fragaria vesca subsp. vesca]
		0	gi|568841693|ref|XP_006474792.1|PREDICTED: probable gal**actin**ol--sucrose galactosyltransferase 1-like [Citrus sinensis] gi|641855238|gb|KDO74032.1| hypothetical protein CISIN_1g004371mg [Citrus sinensis]
		0	gi|657951550|ref|XP_008353333.1|PREDICTED: probable gal**actin**ol--sucrose galactosyltransferase 1 [Malus domestica]
		7.19046e-097	gi|303325104|pdb|3NJ0|AChain A, X-Ray Crystal Structure Of The Pyl2-Pyrab**actin** A Complex gi|303325105|pdb|3NJ0|B Chain B, X-Ray Crystal Structure Of The Pyl2-Pyrab**actin** A Complex gi|303325106|pdb|3NJ0|C Chain C, X-Ray Crystal Structure Of The Pyl2-Pyrab**actin** A Complex

図 4　検索結果表示画面

3.5 コンティグ詳細ページ

各検索結果ページのコンティグ名にはコンティグ詳細ページへのリンクが設定されており，リンク先では該当コンティグに関連するデータをまとめて閲覧することが可能になっている。本ページではGO に関しては関連付けられた全ての GO 情報を表示し，GO ID，GO term およびそれらが属する namespace を示す。blastX 結果も同様に nr データベースに対する上位 20 位までの検索結果を表示し，それはコンティグと相同性が見つかったアミノ酸配列名と e-value からなる。また，本ページからも fasta 形式によるコンティグの塩基配列の取得が可能である。



Output selected sequence format: ◉ fasta ○ contig name list

sequence name:
☑ c111656_g1_i1

blastn database:A_all_contig

	blastn hit contig A_all	blast hit result					
		evalue	identity(%)	query_start	query_end	seq_start	seq_end
☑	A_all_c48479_g1_i1	0	100	300	718	142	560
		7e-063	97.2	93	235	1	143

blastn database:J_all_contig

	blastn hit contig J_all	blast hit result					
		evalue	identity(%)	query_start	query_end	seq_start	seq_end
☑	J_all_c47413_g1_i1	0	100	475	930	1	456
☑	J_all_c74215_g1_i1	0	100	1	417	1	417

Gene Ontlogy

GO ID	Namespace	definition
GO:0006139	biological_process	nucleobase-containing compound metabolic process
GO:0007165	biological_process	signal transduction
GO:0007275	biological_process	multicellular organismal development
GO:0009058	biological_process	biosynthetic process
GO:0009606	biological_process	tropism
GO:0009628	biological_process	response to abiotic stimulus
GO:0009719	biological_process	response to endogenous stimulus
GO:0005634	cellular_component	nucleus
GO:0003700	molecular_function	sequence-specific DNA binding transcription factor activity
GO:0005515	molecular_function	protein binding

blastx databse:nr

blastx hit gene	evalue
gi|297834672|ref|XP_002885218.1|predicted protein [Arabidopsis lyrata subsp. lyrata]	1.28463e-074
gi|297331058|gb|EFH61477.1| predicted protein [Arabidopsis lyrata subsp. lyrata]	
gi|15229074|ref|NP_188387.1|auxin-responsive protein IAA31 [Arabidopsis thaliana] gi|46395870|sp|Q8H174.2|IAA31_ARATH RecName: Full=Auxin-responsive protein IAA31; AltName: Full=Indoleacetic acid-induced protein 31 [Arabidopsis thaliana] gi|9294148|dbj|BAB02050.1| unnamed protein product [Arabidopsis thaliana] gi|15810012|gb|AAL06933.1| AT3g17600/MKP6_15 [Arabidopsis thaliana] gi|49616379|gb|AAT67086.1| IAA31 [Arabidopsis thaliana] gi|298108635|gb|ADB93680.2| indole-3-acetic acid inducible 31 [Arabidopsis thaliana] gi|298108639|gb|ADB93682.2| indole-3-acetic acid inducible 31 [Arabidopsis thaliana] gi|298108641|gb|ADB93683.2| indole-3-acetic acid inducible 31 [Arabidopsis thaliana] gi|304322703|gb|ADL70804.1| indole-3-acetic acid inducible 31 [Arabidopsis thaliana] gi|304322705|gb|ADL70805.1| indole-3-acetic acid inducible 31 [Arabidopsis thaliana] gi|304322707|gb|ADL70806.1| indole-3-acetic acid inducible 31 [Arabidopsis thaliana]	3.7377e-074

図 5　コンティグ詳細ページ画面

4. 今後の展開

本データベースは形式の異なるデータをそれぞれ別のテーブルとして取り込み，閲覧する場合には Perl による各テーブル間の結合と動的な HTML ページの作成を行っている。そのため，新たに様々

な形式のテーブルを取り込みデータベースを拡充することが可能である。また，動的表示であることは，ページ作成を行っている Perl スクリプトの変更によってデータベースの形式を変更することなく，出力するページの形式変更および追加されたテーブルの情報を表示できることを示している。現在までに *R. aquatica* の気中条件での茎頂サンプルによる RNA-seq 発現情報解析が行われ，データベースに組み込まれている。今後の計画として，茎頂以外のサンプルを用いた RNA-seq や水中条件での RNA-seq が進行中である。また，de novo アセンブルによるゲノム配列の取得も行われている。このように様々な方面から *R. aquatica* のゲノムおよび遺伝子情報が解析され，多くのデータが得られてきている。今後も取得された多様な解析結果をデータベースとして登録することで各研究間の連携を強め，多角的な視点からの知見を得ることで *R. aquatica* を網羅的に解析していきたいと考えている。

謝辞

本研究は，京都産業大学第 3 次総合研究支援制度新規研究課題挑戦支援プログラム「葉の形態の表現型可塑性のメカニズムと進化過程の解明」(課題番号 E1504) の研究の一部として実施された。また，研究の一部は，平成 27 年度私立大学戦略的研究基盤形成支援事業「植物における生態進化発生学研究拠点の形成―統合オミックス解析による展開―（課題番号 S1511023)」，および，平成 24 年度科学研究費助成事業（学術研究助成基金助成金（若手研究（B))「葉の形態の表現型可塑性の分子基盤の解明:環境に応じて葉形を変化させる植物の研究 (課題番号 24770047)」) の支援を受けて実施した。

参考文献

1, The Gene Ontology Consortium. Gene Ontology Consortium: going forward. (2015) Nucl Acids Res 43 Database issue D1049-D1056.

Construction of a genomic and transcriptomic database for *Rorippa aquatica*

Tomoaki SAKAMOTO
Seisuke KIMURA

Abstract

Rorippa aquatica shows several interesting traits, such as leaf formation in response to environmental conditions and regeneration from a piece of leaf in the absence of exogenous plant hormones. However, complete genome and transcript data for this species are not yet available and several ongoing analyses are underway to obtain these using next-generation sequencing. At this stage, the obtained datasets are quite large and the outputs are in different formats.

In this study, we tried to construct a database to store and integrate data from various studies on *R. aquatica*. Furthermore, to improve the accessibility of these data, we have also developed a graphical user interface.

Keywords : *Rorippa aquatica*, *de novo* assembly, database, SQLite, Perl/CGI