第1章 第1節

果樹研究のためのゲノム情報

農研機構果樹研究所 カンキツ研究領域 藤井 浩

カンキツやリンゴ、ブドウ、モモといった主要な果樹を含めて、ウェブ 上のサーバから果樹研究に利用可能なドラフトゲノム情報が公開されている (表1-1-1).これら情報を利用することで、研究対象とする遺伝子ホモログ (homolog)のゲノム上の個数や位置情報、上流域・下流域の塩基配列やエキ ソン-イントロンの構造を簡単に入手できる.また、全ての予測遺伝子の塩基 配列もマルチFASTA形式のファイルとしてサーバに保存されているので、ダ ウンロードして独自のマイクロアレイの設計に活用できる.実験計画をたてる 際に、既知のゲノム情報を参照することで、研究の効率化が期待できる.本節 では、ウエットな実験を中心に研究を進めている研究者を対象に、果樹研究に とって基本的なゲノム情報をウェブブラウザを使って入手する方法と、入手し たゲノム情報をおもにWindows環境で閲覧する方法を解説する.紹介するソ フトウエアには、有償のものを含め、できるかぎりグラフィカルなインターフ ェースやサポートを備えているものを選んだ.

本格的なLinuxを利用したバイオインフォマティクス用の環境の構築につい ては、巻末に参考図書を掲げるにとどめて、本節では述べない.次世代シーク エンサに関するバイオインフォマティクスツールに関しては次節で、ゲノムア ノテーション (genome annotation) に関するツールは本章第3節で、ゲノム データからDNA多型 (polymorphism) を検出するためのツールについては本 章第4節で、マイクロアレイ設計に関するツールは本章第5節で詳述する.

1) Phytozome

ゲノム情報は、多くの場合、それぞれのゲノム解読プロジェクトのサイト から公開されているが(表1-1-1),米国エネルギー省の研究所であるToint Genome Institute (JGI) が開設しているゲノムサイトPhytozome (URL1-1-1. Neupane, R. et al. 2012) には、2016年1月時点で、57種の生物種に関する 63のゲノムデータベースが収録されている (図1-1-1). この中には、クレメ ンティン (Citrus clementina v1.1), スイートオレンジ (Citrus sinensis v1.1), リンゴ (Malus domestica v1.0), ヨーロッパブドウ (Vitis vinifera Genoscope 12X), モモ (Prunus persica v1.0, v2.1) など主要な果樹とモデル植物であるシ ロイヌナズナ (Arabidopsis thaliana TAIR10) などのゲノム情報が収録されて いる. (学名の後ろの番号はそれぞれのゲノムアセンブリのバージョンを示し ている). Phytozomeの利点は、共通の操作方法でゲノム情報が利用可能で、 複数種を指定した横断検索もできるところにある。なお、果樹の公開ゲノム 情報のうち、中国のグループが発表しているスイートオレンジ(Citrus sinensis Annotation project. URL 1-1-2) やチュウゴクナシ (Pvrus bretschneideri: Pear Genome Project, URL1-1-3)のゲノム情報はphytozomeに収録されて いない. また, バラ科ゲノムデータベース (Genome Database for Rosaceae. GDR, URL1-1-4) に公開されているモモのマーカー情報 (図1-1-2) のように, 各ゲノムプロジェクト独自の解析については、収録されていない、こうした独 自の内容の解析データが追加・更新されることもあるので、各ゲノムプロジェ クトのオリジナルサイトも巡回する必要がある.

(1) 遺伝子検索

研究対象とする遺伝子ホモログをゲノム情報から網羅的に取得する場合, 遺伝子名によるキーワード検索と配列の類似性によるBLAST (URL1-1-5, Altschul *et al.* 1997)検索の双方の検索を必ず行うべきである。キーワード 検索のみの場合,配列が類似しているホモログを収集できない可能性があり,

名称·学名·系統	ゲノムデータベース	URL番号 (URL目次を参照)	アセンブリバージョン	紫文
カレメンティン Citrus clementina Clemenules (haploid)	Citrus Genome Database Phytozome	URL1-4-6 URL1-1-1	v0.9, v1.0 v1.0	Wu <i>et al.</i> 2014
スイートオレンジ <i>Citrus sinensis</i> Ridge Pineapple	Citrus Genome Database Phytozome	URL1-4-6 URL1-1-1	v1.0 v.1.1	Wu <i>et al.</i> 2014
スイートオレンジ <i>Citrus sinensis</i> Valencia	Citrus sinensis Annotation Project	URL1-1-2	v1.0, v2.0, v2.1	Xu <i>et al.</i> 2013
リンゴ Malus domestica Golden Delicious	Genome Database for Rosaceae (GDR) Phytozome	URL1-1-4 URL1-1-1	v1.0, v1.0p, v2.0, v3.0.a1 v1.0	Velasco <i>et al.</i> 2010
ヨーロッパブドウ Vitis vinifera PN40024	Grape Genome Browser (Genoscope) Phytozome VvGDB	URL2-4-1 URL1-1-1 URL1-4-5	8X, 12X 12X 12X	Jaillon <i>et al.</i> 2007
푼푼 <i>Prunus persica</i> Lovell (double haploid)	Genome Database for Rosaceae (GDR) Phytozome	URL1-1-4 URL1-1-1	v1.0, v2.0.a1 v1.0, v2.1	The International Peach Genome Initiative 2013
セイヨウナシ <i>Pyrus communis</i> Bartlett	Genome Database for Rosaceae (GDR)	URL1-1-4	v1.0	Chagné <i>et al.</i> 2014
チュウゴクナシ Pyrus bretschneideri Dangshansuli	Pear Genome Project	URL1-1-3	v1.0	Wang <i>et al.</i> 2013

表1-1-1 温帯果樹のゲノムデータベース

第1章第1節 果樹のゲノム情報

3

Amaranthus hypochondriacus v1.0	Eucalyptus grandis v2.0	Physcomitrella patens v3.3
Amborella trichopoda v1.0	Eutrema salsugineum v1.0	Populus trichocarpa v3.0
Ananas comosus v3	Fragaria vesca v1.1	Prunus persica v1.0
Aquilegia coerulea v1.1	Glycine max Wm82.a2.v1	Prunus persica v2.1
Arabidopsis halleri v1.1	Gossypium raimondii v2.1	Ricinus communis v0.1
Arabidopsis lyrata v1.0	Kalanchoe marnieriana v1.0	Salix purpurea v1.0
Arabidopsis thaliana TAIR10	Linum usitatissimum v1.0	Selaginella moellendorffii v1.0
Boechera stricta v1.2	Malus domestica v1.0	Setaria italica v2.1
Brachypodium distachyon v2.1	Manihot esculenta v4.1	Setaria italica v2.2
Brachypodium distachyon v3.1	Manihot esculenta v6.1	Setaria viridis v1.1
Brachypodium stacei v1.1	Medicago truncatula Mt4.0v1	Solanum lycopersicum ITAG2.3
Brassica rapa FPsc v1.3	Micromonas pusilla CCMP1545 v3.0	Solanum tuberosum v3.4
Capsella grandiflora v1.1	Micromonas sp. RCC299 v3.0	Sorghum bicolor v2.1
Capsella rubella v1.0	Mimulus guttatus v2.0	Sorghum bicolor v3.1
Carica papaya ASGPBv0.4	Musa acuminata v1	Sphagnum fallax v0.5
Chlamydomonas reinhardtii v5.5	Oryza sativa v7.0	Spirodela polyrhiza v2
Citrus clementina v1.0	Ostreococcus lucimarinus v2.0	Theobroma cacao v1.1
Citrus sinensis v1.1	Panicum hallii v2.0	Triticum aestivum v2.2
Coccomyxa subellipsoidea C-169 v2.0	Panicum virgatum v1.1	Vitis vinifera Genoscope.12X
Cucumis sativus v1.0	Phaseolus vulgaris v1.0	Volvox carteri v2.0
Eucalyptus grandis v1.1	Physcomitrella patens v3.0	Zea mays 6a

図1-1-1 Phytozomeから提供されているゲノムデータ 57種の生物について63種類のゲノムデータが収録されている

GDR Genome Database							March	Bearsh Login Home Contact	
General	16.9	Species	Data	bearch	Tools	Breeders Toolbox	Community		¥
Prunu	is pers	ica Whol	e Gend	me v1.0	Assen	nbly & Annotati	on		
Ma The Dow	Markers The Phone persise v1.0 genome methers likes are available in FASTA and Excel format with links to Officence. Downlineds						it: Olivouse	Resources • View pseudomaleculas in Gilcowse • Syntemy with streactery and apple • Details • Details	
Phy	Phones genetic metion sequences in peach (FASTA Re) Phones_mep_markers_sequences leafs						markers_sequences.lests	Assembly	
Prunus genetic marker sequences in peach: SSRs only (FASTA file) Prunus_map_markers_					Rs only (Fi	ASTA file)	markers_sequences_BSRs.tasta	Ratata - Assembly Referencets	
Pronus genetic markens (Encel) GOR_markens_ptrowse.vls						_gbrowse.xis	+ Hamology		
ResOOD Manans aligned to periome (OPP3)					RosCos_vs_P	teachumarkans.gfD.gz	Matters ShPa Functional Analysis		

図1-1-2 GDRのモモの遺伝マーカーのダウンロード画面

BLAST検索のみだと配列が類似していないが機能が類似しているホモログを 収集できないからである. Phytozomeのトップページ上部のメニューバーの 左から2つ目の「Tools」を選択すると(図1-1-3), プルダウンメニューか ら「Keyword search」,「BLAST」,「BLAT (URL1-1-6. Kent 2002)」など のツールが選べる.例としてクレメンティンを対象に「MADS」という名称 の遺伝子を検索してみる.「Keyword search」を選択して、「Citrus clementine v1.0」を選択したのち、キーワード入力欄に「MADS」と記入して(図1-1 -4)検索を実行すると、図1-1-5のように遺伝子の注釈(アノテーション、 annotation) に「MADS」という文字列が含まれている遺伝子が列挙される. 目的のMADS遺伝子の行の左端の白抜きの「B|を選ぶとゲノムブラウザが起 動して、当該遺伝子のゲノムにおける位置や周辺の遺伝子の状況が画像として 表示され (図1-1-6), 「G」を選ぶとPANTHER (Protein ANalysis THrough Evolutionary Relationships, URL1-1-7, Thomas et al. 2003) ♦Pfam (URL1-1-8), KOG (EuKarvotic Orthologous Groups, URL1-1-9), 「GO」(遺伝子) オントロジー, gene ontology, URL1-1-10, Ashburner et al. 2000) といった データベースへのリンクリストが表示され(図1-1-7),詳しい遺伝子機能情 報を入手できる.

(2) 遺伝子構造の閲覧と上流域塩基配列の取得

検索した遺伝子の遺伝子構造を閲覧するための方法を説明する.図1-1-7の中段付近にある「Sequence」タブを選択し、白抜きで現れる「Genomic sequence」ボタンを選択すると、口絵1-1-1のように、5'UTRが緑色に、 CDSが青色に、3'UTRが桃色に、構造別に塗り分けられて表示される.また、 この遺伝子の上流域3,000塩基を取得するには、口絵1-1-1の塩基配列の上部 にある「upstream」の表示の右隣の入力欄に「3000」と記入して、右側の「Submit」 ボタンを選択すると、図1-1-8のように上流域の3,000塩基の塩基配列が表示 される.



図1-1-3 Phytozomeのトップページ

Toolsメニューを開いた状態. Toolsメニューに「KeyWord search」,「BLAST」,「BLAT」などがある.「BLAT」は質問配列がゲノム配列中のどの部分にヒットするかゲノムランディング (genome landing) 探索するのに用いる.



図1-1-4 Phytozomeにおけるキーワード検索画面

ここでは, Citrus clementinaが選ばれ,「MADS」をキーワードとしている.



図1-1-5 Phytozomeにおけるキーワード検索結果表示 アノテーションに「MADS」を含む遺伝子が表示される.

JGI X Phytozo	meA# 10.3	20	HOME			
Available Tracks	Citrus clemer	ttina v1.0 - File VS	ev tep		00 Stee	Falleren
X filter by test	0	8,000,000	10,000,000	15,000,000	20,000,000	25,000,00
Cape Reference sequence User Blact Results	2		Q Q Q Q	cartuid_6[+]+United_620831401.3 20.46	156710760 (H4.3 Go) 1.500	21.875,000
* Alignments BUITX Plant Proteins BUITX Plant Proteins BUITX Plant Proteins Plant Aligned STruction Plant Aligned Struction Plant Aligned	A BOTH	erritoriastism Ga Ellivrootise newoniet	evitüri 162m Cole	Cost octation	Ge	pr1012564m
Transcripts P Attenuative Transcript Contact. Decision Accessibility / Section SNI Statement development of the University of 5	x				@ENERG	Y Office of Science

図1-1-6 Phyozomeのゲノムブラウザ表示 ゲノム上での遺伝子の位置関係がわかる. 拡大縮小が可能である.

(3) ゲノム情報のダウンロード

ゲノムデータを取得するには、Phytozomeのアカウントを取得してログイン する必要がある.ログインすると,検索結果を保存することもできる.一般に、 ゲノムデータベースを使用する際は、アカウントを取得して使用すると利便性 が高まる.ログイン後、図1-1-3のメニューバーの「Downlaod」を選択する. Phytozomeのバージョンを選択するプルダウンメニュー表示されるので、と くに理由がなければ最新のバージョンを選択する.すると、「Genome Portal」 のページが表示される(図1-1-9).種名が表示されているフォルダの中から、



図1-1-7 Phytozomeの遺伝子アノテーション表示 画面下部に示されているようにPANTHERやPfam, GOのアノテーションを閲覧できる.

例えば「Cclementina」のフォルダアイコンを開くと、図1-1-10のようにダウ ンロード可能なファイルが表示される.ダウンロードしたいファイルにチェッ クを入れて緑色のボタン「Download Selected Files」(図1-1-9)を選択する とダウンロードが開始される.

図1-1-10に示されている「annotation」フォルダには、coding sequence (CDS) 配列やタンパク質配列のファイルがマルチFASTA形式で納められて おり、「assembly」フォルダにはゲノム配列としてのスキャフォルド(scaffold) がマルチFASTA形式で納められている.ファイル名の拡張子のうち、faは FASTAファイルを示し、gzはUNIX互換OSでよく用いられるファイル圧縮 ソフトGZIPによって圧縮されたファイルであることを示す.Windows環境 でもLhaplusやLhacaなどのフリーソフトウエアを用いてgz形式で圧縮され たファイルを解凍することができる.また拡張子がgffで示されるファイル はgeneral feature format version3 (GFF3)の意味で遺伝子構造情報が格納



図1-1-8 Phytozomeの遺伝子上流域の表示 「upsteam」の欄に「30000」を入れると上流域の3,000塩基が表示される

第1章 ゲノム情報とバイオインフォマティクスツール

I JGI 凝	Genom	e Portal		JELHOME SENOME PORTAL MO HIROSHI FUJII (HFUJI)(BAFFRCGD.	VPORTAL LOGOUT	
						PhytozomeV
NO DOWNLOA	D. PHOTOSOME	BLAST STATUS	HELP			
Download via Gi described here	obus) - to downlo	ad large files (m	nore than 100MB	3) or multiple files we recommend	dusing <u>Globus serv</u>	ice
Some of the data av science. By accessi publication by JCI at publications (a.g. re arbiter regarding pu Each directory with Download Selec	valiable on this portal a ing these data, you ag ing the clabostons of smaller for each orga served analyses) of th blioation plans. Scient in this portal has a file ted Files Expand?	In pre-publication de ree not to publish any lis comprehensive go rism. During this wai e JGI and collaborati lists are strongly enco clearribing the release NI (CollapseAI)	Its and may contain a y articles containing a anome analysis. The ting period, the data tors. A principal collab ouraged to contact the us policy of that data.	entres. The goal of cur pooky is that early we analyses of grants or generatic date on a wh are mitrictions will be illed on the publication foreign of the analysis of publication increater of thempion," listed in the organic principal collaboration and JSB abust their People who access data are expected to a	rease should enable the p nole genome or chromosor on of the whole genome de that does not compete din the finite page and is the po- intentions and any potent othere to those policies.	rogress of he scale prior to scatprion or as scatp with planne int of contact an al collaboration.
Disease in an in min	nd that downloading ta	ape tiles (🖾) can	take a few minutes.		Lesend	
maga wep in me						



8 🗇 🛛 Colemen	514
* CD (2 anno	tardion
- 28	/huturomeV10/ Columentine, 162, v1.4 annotation, Jolluter; 5 MB; Sun Jan 12 17:45:56 PGT 2014 Caute Expires; Wed Feb 03 06:05:44 PGT 2016 👔 📾
- 20	PhytocomeV10 Columenting 162 v1.4 eds.fs.gr 10 MB; Bun Jan 12 17 08:32 PST 2014 Cache Expire: Web Feb 03 08:05 44 PST 2016 👔 📾
- 20	PhotocomeV10: Golementine, 162, v1.4.oda, primary/Transcript/Inty/Augr. B MB; Sun Jan 12 17.08.31 PST 2014. Coche Expires: Wed Feb 03.08.05.44 PST 2016 👔 🛲
- 20	Photosome/YID Columenting 162 v1.4 gene.gf2.gg 3 MS; Sun Jan 12 17:08:27 PST 2014 Capte Expine: Wed Feb 03:08:05:44 PST 2016 👔 📾
- 20	Photocome/TiD: Columentine: 162: v1.4 gene.exons.gdD.gr 5 MB; Sun Jan 12 17:08:33 PST 2014 Code Expires: Wed Feb 03 08:06:44 PST 2015 🕕 📾
- 20	Phylosome/YiD: Sciementine, 162, v1.6 protein.fe.gr 6 MB; Sun Jan 12 17:08:32 PST 2014 Cache Expires: Wed Feb 05 08:06:44 PST 2016 🕕 🖷
- 2 🗖	PhotoconeV10: Colementine, 162, v1.6 protein, primary/Tensorio/Only/le.gz 6 MB; Sun Jan 12 17:08:31 PST 2014 Cache Expiner: Wed Feb 05 08:06:44 PST 2015 📋 📻
- 20	Physiciane/Y10 Columentine. 182 v1.4 representative assembly v1.973.gr 6.405. Sun Jan 12 17:08:28 PDT 2014 Cestre Expires: Wed Feb 03 08:06:44 PDT 2016 👔 ன
- 20	Phytosane/Y10: Columentine, 162, v1.6 transcript, fa.gz 13 MB; Sun Jan 12 17:08:29 PST 2014 Coche Expine: Wed Feb 03 08:05:44 PST 2016 🚺 🗃
- 20	PhotocomeV10: Columentine: 152 v1.4 transcript primeryTranscriptOn(x.fs.pg: 12 NI); Sun Jan 12 17:08:30 PST 2014 Cache Expires: Wed Feb 03-08:05:44 PST 2016 👔 🖨
8 (D (2 anna)	mbly
- 20	Phylazane/Y10: Galementina, 162, v1.fa.gz, 63 MB; Sun Jan 12 17:06:36 PST 2014 Cache Expines: Wed Feb 63 06:05:44 PST 2016 👔 📷
- 20	PhotocomeVTD: Columentine: 152: v1.handmasked.fa.gz; 61 MB; Sun Jan 12: 17:05:45 PST 2014 Cache Expine: Wed Feb 03:06:05:44 PST 2015 👔 📾
	PhotocomeV10: Columenting: 162 v1.coftmasked Augr: 87 MB; Sun Jan 12 17:08:39 PST 2014 Cache Expires: Wed Feb 03 08:06:44 PST 2016 🚺 🚓
200	utzome/10: Opiementina, 162 v1.8.medne.bd 4 KB; Pri Sep 18 10.55:36 POT 2015 Cache Expires: Sat Sep 17 16:05:39 POT 2016 🔲 🛶
- 20 0	vezzonev/12 DataRevasePolicy.Attel 1 KR; Wed Aug 13 18:29:38 PDT 2014 Cache Explore: Tuo Jan 05 17:01:41 PST 2015 🕢 🚾

図1-1-10 Phytozomeのダウンロードファイルの例 クレメンティンのゲノムデータファイルを示している.

されていることを示す.図1-1-6に示したゲノムブラウザの画像は、GFF3 ファイルのデータを基づいてグラフィカルに遺伝子構造を表示した結果であ る.テキストデータとしてGFF3ファイルを閲覧する方法については後述する. repeatmaskedなどmaskedが付されたファイルは、繰り返し配列がNでマスク された配列であることを示す.繰り返し配列があるとアラインメントや類似 性検索が正しく行えないことがある.それぞれの詳しいファイルの内容につい ては,同じくダウンロードサイトにあるCclementina_182_v1.0. readme. txt や Data_Release_Policy. htmに記述されている.

2) TAIR

研究対象とする遺伝子ホモログについて、ゲノム情報を取得する方法の一 つとして、モデル植物のシロイヌナズナ(Arabidopsis thaliana)のゲノムアノテ ーションデータベースであるTAIR(The Arabidopsis Information Resource, URL1-1-11, Huala et al. 2001)の利用がある. TAIRでは、各遺伝子にGOを はじめ、発現情報や文献情報など豊富なアノテーションが付されているので、 当該遺伝子に関する既往の情報を広範に得ることができる. Phytozomeの場合 と同様に、研究対象とする遺伝子ホモログを網羅的に取得する場合、遺伝子名 によるキーワード検索と配列の類似性によるBLAST検索の双方の検索を必ず 行うべきである.

3) Taxonomy Browser

アメリカ国立衛生研究所 (National Institutes of Health; NIH)の所属機 関である国立生物工学情報センター (National Center for Biotechnology Information, NCBI) が管理・運営している国際塩基配列データベースの一つ であるTaxonomy Browser (URL1-1-12) を利用すると,研究対象とする果 樹の属や種に関して,NCBIの統合データベースEntrez (URL1-1-13) に登 録されている塩基配列情報や遺伝子発現情報などを網羅的に取得することがで きる.例として,Taxonomy Browserの検索欄に「citrus」をキーワードにし て検索すると,カンキツ属 (*Citrus*)の種ごとに整理されたEntrezのポータル ページへのリンクが表示される (図1-1-11).「Citrus clementina」を選択す ると,右端の表に*Citrus clementina*に関するEntrezのすべての登録情報の数が 表示される (図1-1-12).右端の表を見ると,この時点での*Citrus clementina* の「Nucleotide」の登録数が43,615であることがわかる.数字「43,615」に付 さいるリンクを選択すると,「Nucleotide」のリストが表示される (図1-1-



図1-1-11 NCBI Taxonomy Browserでの検索結果

キーワード「Citrus」で検索した.上部の「Citrus」をクリックすると, *Citrus*属全体の Entrezのデータを観ることが出来る.



図1-1-12 NCBI Taxonomy Browserでの検索結果表示

右側の表にクレメンティンに関してEntrezに収録されている各種データの数が表示される.

13). 43,615の「Nucleotide」をFASTA形式のファイルでダウンロードするには、図1-1-13の右上方の「send to」を選択して現れたプルダウンメニューの「Choose Destination」で「File」を選択し、「Format」で「FASTA」を選択して、「Create File」をクリックすれると、43,615の「Nucleotide」の塩基配列がFASTA形式のファイルとしてユーザのPCにダウンロードされる.ダウンロードした塩基配列は、自身のPC内(ローカル、local)で実行するBLASTのデータベース作成用などに利用できる.ローカル環境でのBLASTの実行方法や、ダウンロードした塩基配列をBLASTのデータベースにする方法は後述する.

また、図1-1-14に示したように、左上方の「Summary」を選択すると 「Format」を選択するプルダウンメニューが現れ、各遺伝子の画面表示のフォ ーマットをGenbank形式などに変更することもできる.

S NCBI Resources C	How To 🕤		Mull My NOB Sign Out		
Nucleotide	Nacionide 0 [bid85681[Organism:ndexp] Create alert Advanced	0)	Eranth Hole		
Bandles Pares (KJ, M10) Customione Miniscule types percent CINARIAN (BJ, R02) Miniscule types Castomione Bource database NBDC (Standinek) (7, M02) Reflec (BJ, S505) Customione	Summary + 20 per page + Sort by Datasit order + Items: 1 to 20 of 43615 Citize 200319 nucleotide sequences. Nucleotide (43010) EST (155800) Citize dementine autivar Ciencerules uncleosed genomic sosfield so 1. stolgan sequence 2.8,940,139 pp lenar DAA Accession NW,000873031 1 de SF7823499 Gentlenit FAS35. Gradhiss	Send to + Filtens: Manage Choose Destination @File OClobated Coolections Download 43015 Items. Format Format FastA 0 Soft ty Othera eres 0 Create File	Films		
Genetic compartments Chicroplast (11) Plastid (11)	 Citrus clementina hypothetical protein (CiCLE_v10006722mg) mRNA 1,555 bp linear mRNA Accession: XM (06455465.1 G): 56702043 Geröfen: RASTA, Graphica 	Search	See more		
Sequence length Custom nange Retesse date Custom nange	Citrus demertine hysothetical anatelin (CiCLE_v1001)0455mg1 mRNA.complete.cds Recent activity TatTo Accession XM, 00453444.1 GE 50403041 Complete.cm Complete.cm				
Revision date Custom range Clear all	 Citrus clementina hysothetical protein (CiCLE_v10008098mg) mRNA 2,272 bp linear mRNA Accession: XX, 0645045.1 Git 95702009 Conflace: 5653. doublins 	, complete cds 🛛 Cinus Q, 6x62706(Cr	Russone Second ganism.exp) (2) Centre		

図1-1-13 NCBI Nucleotideの検索結果表示1

クレメンティンの43,615の登録NucleotideがSummary形式で表示されている.右上の「Send to」を選択するとダウンロードの方法とファイル形式のプルダウンメニューが表示される.

S NCBI Resources	s⊠ How To ⊡		Muli My NOBI 5	Sign Out
Nucleotide	Nuclearity 0 bid Creat	85681[Organism:noexp] ta alant: Advanced	O Seath	Help
Species	Summary + 20 per p	ege = Sort by Default order = Send to: =	Filters: Manage Filters	
Parma (FAUTE) Format Datamitas - Sammany Molecule types Gandbark (KI) PARTA (SL) Gandbark (KI) PARTA (SL) Gandbark (KI) PARTA (SL) Catamitas - ASN 1 Datamitas - ASN 1		615	Find related data Database: Select 0) Find bars	۲
Bource databases INSDC (Gentiank) (7,0 FlatSeg (35,955) Customize	Accession List Of List Gentians ENSU	er DNA SS3301 1 Gr. 667323469 A Gradica	Search details txid85681(Organian: noexp)	6
Genetic compartmenta Chicropiaet (11) Plantid (11)	Citzus clementin 2. 1,555 bp linear Accession: XM_00 DenBaria FRSD	na hypothetical protein (CICLE_v1008/722mg) mRNA, complete cds mRNA 54634631 01 56782043 A. Oroshina	Bearth 5a	
Sequence length Custom range Release date Custom range	Citrus clementio 3. 1,347 bp linear Accession XM_X Destians (ASI)	n Hopochetikasi proleinin (CICLE_v10010435mg) mRHA, pomoletie oda mRMA SHESHAA 0. 607823041 A. Granbia	Recent activity TumO Q. biot5551(Organismunoeup) (4361	er same Ro
Revision date Custom range Clear.al	Citrus clementir 4. 2,272 bp linear Accession: XM_00 Gentiaris EASIO		E Cinus Q. bak(3708(Organismum) (2)	Serone

図1-1-14 図1-1-13NCBI Nucleotideの検索結果表示2

右上のSummaryを選択すると、表示形式をプルダウンメニューから選ぶことができる.

4) データ閲覧用のソフトウエア

次世代シークエンサのデータ解析に用いるような本格的なバイオインフォマ ティクスツールついては、第1章第2節で紹介する.本節では、上記で得たゲ ノムデータを利用するための汎用的ソフトウエアを紹介する.多くの場合、ゲ ノムデータのファイルサイズは大きいので、Microsoft社のWordやExcelでの 閲覧は難しく、下記に紹介するようなソフトウエアが実用的である.

(1) テキストエディタ

Mery (URL1-1-14) やMKeditor for Widows (URL1-1-15), 秀丸エディ タ (URL1-1-16) といったテキストエディタの多くは,本来はコンピュータ プログラムの記述を目的としたツールであるが,大量の塩基配列データのよう なファイルサイズの大きいテキストデータを閲覧するのに適している.また, テキストエディタの多くは,正規表現による文字以外の制御記号の編集が可能 である.例えば,タブ記号を改行記号に置換がすることができるので,ファイ ル形式の変換にも役立つ.grepという機能を用いると特定の文字列を含んだ 行をすべて調べることができるので、複数のマルチFASTAファイルにおける 特定のアノテーションが含まれる行の書き出しが容易である.テキストエディ タを選択する場合、ソフトウエア流通サイトのVector(URL1-1-17)を利用 すると、多数のフリーソフトウエアやシェアウエアの中から好みのソフトウエ アを探すことができる.巨大なテキストファイルを閲覧することができるGiga Text Viewer(URL1-1-18)や前述のファイルを圧縮・解凍するためのアー カイブ・ユーティリティなども紹介されている.

(2) データベースソフトウエア

前述のGFF3ファイルを閲覧する場合は、有償のデータベース作成ソフト ウエアであるFileMaker Pro (FileMaker社, URL1-1-19)の利用が便利であ る.GFF3ファイルのサイズは非常に大きいことが多いが、GFF3ファイルを FileMakerのショートカットアイコンにドラッグ・アンド・ドロップするだ けで、GFF3ファイルの内容が表示され、簡易なデータベースとして利用でき る.同様に、(commma separated value, CSV)ファイルやExcelファイルも ドラッグアンドドロップだけでFileMaker形式のファイルに変換できる.また、 FileMakerから、CSVやExcel形式のファイルに出力することも容易である. 簡易なリレーショナル・データベースを構築できるので、複数のデータセット 間のリレーションを取って、任意のデータの組み合わせでの表示や出力が可能 である.FileMakerは人気のあるソフトウエアなので、わかり易い参考書も複 数出版されていて、自学も容易である.巻末の参考図書に、その一部を掲げた.

5) ローカルBLASTおよびファイル形式変換のためのソフトウエア

有償ソフトウエアのGENETYXパッケージ(株式会社ゼネティクス, URL1 -1-20)やCLC Main Workbench (QIAGEN社, URL1-1-21)などを利用す ると,自分用のBLASTデータベースを作成して,ローカル環境でBLASTを実 行できる.例えば,自身が解読したゲノム塩基配列などをデータベース側とし て、BLASTが実行できるようになる.また、上記3)のTaxonmy Browserか らダウンロードした対象樹種の遺伝子のタンパク質配列をデータベースにする ことで、ノイズの少ないBLASTの結果が得られる.

また、これらの塩基配列・タンパク質配列処理ソフトウエアはファイル形式 の変換にも利用できる.バイオインフォマティクス分野では表1-1-2のような 多くのファイル形式が使用されるが、CLC Main Workbenchは、これらのフ ァイル形式の入出力ができるので、例えば、マルチFASTA形式の塩基配列を 入力して、CSV形式で出力することも可能である.CLC Main Workbenchは 第1章第2節で解説するCLC Genomics Workbenchから、次世代シークエン サデータのアセンブリ機能を除いたソフトウエアである.

6) もうすこし先のバイオインフォマティクス

バイオインフォマティクス手法を利用するには、Apple社のMacは比較的有 用である. MacのOSであるMac OSXはUNIX系のOSであるので、ターミナル というアプリケーションを使うとUNIXコマンドを使用できる. UNIXコマン ドを用いると、例えば、多数のファイルの結合や名前の付け替え、ファイル サイズの大きなテキストファイルの閲覧も容易である. Macによるバイオイン フォマティクス解析の方法を解説した書籍を巻末の参考図書に掲げた. なお、 Macに備えられているBoot Campというソフトウエア用いるとMac OSX環境 とWindows環境を併存させることができるので、MacとしてもWindowsマシ ンとしても使用できる. また、VMware Fusion (VMware社, URL1-1-22) を使用すると、Mac上でWindowsを実行することもできる.

日常使用しているWindowsやMacでは困難な大規模計算を行う場合,一つ の方法として,フリーの統計解析ソフトR(URL1-1-23)の利用がある.Rに はBioconductor(URL1-1-24)というゲノムデータ解析用のパッケージがあ り,プログラミングをする必要は少ない.Rは,研究機関や大学の共用の計算 用サーバに導入されていることが多いので,こうしたサーバのアカウントを入 手できれば,手許のPCから計算用サーバのRを操作して,サーバに大規模な計

ファイル名	拡張子	ファイルの種類
ACE	.ace	コンティグ
Phylip Alignment	.phy	アラインメント
GCG Alignment	.msf	アラインメント
Clustal Alignment	.aln	アラインメント
Newick	.nwk	ツリー記述
FASTA	.fsa/.fasta	酉己歹」
GenBank	.gbk/.gb/.gp	配列
GCG	.gcg	酉己歹」
PIR	.pir	配列
Staden	.sdn	配列
DNAstrider	.str/.strider	酉己 <i>歹</i> 」
Swiss-Prot	.swp	タンパク質配列
Lasergene	.pro	タンパク質配列
Lasergene	.seq	塩基配列
Embl	.embl	塩基配列
Nexus	.nxs/.nexus	配列, アラインメント, ツリー記述,他
CLC	.clc	配列, アラインメント, ツリー記述,他
Text	.txt	テキスト
CSV	.csv	カンマ区切りテキスト
ABI	.abi	塩基配列トレースファイル
AB1	.abl	塩基配列トレースファイル
SCF2	.scf	塩基配列トレースファイル
SCF3	.scf	塩基配列トレースファイル
Phred	.phd	塩基配列トレースファイル
mmCIF	.cif	構造
PDB	.pdb	構造
BLAST Databese	.phr/.nhr	BLASTデータベース
Vector NTI Database	-	酉己歹 」
Vector NTI	.ma4/.pa4/.oa4	酉己歹」
Gene Construction Kit	.gcc	配列

表1-1-2 バイオインフォマティクスで用いられるファイル形式

算を行わせることは、それほど難しくない.とくに繰り返し計算を行う場合に、 ループ機能が使えるので便利である.研究機関や大学によっては、グラフィカ ルなインターフェースで計算用サーバのRを使用する環境であるR studioを備 えている.RはWindowsでもMacでもインストールできるので、ローカルな環 境で利用することもできる.Rに関する解説書を巻末の参考図書に掲げた.こ のほか、外部のサーバでインフォマティクス解析を行う方法として、Galaxy (URL1-1-25, Giardine, B. *et al.*' 2005) やGalxy/NIAS (URL1-1-26), DDBJ Annotation Pipeline (URL1-1-27, Kaminuma, E. *et al.*' 2010) などがある. こ れらについては, 第1章第2節や第4節でも紹介する.

バイオインフォマティクス関連のツールやデータベースは、すべてではない が、バイオサイエンスデータベースセンター(National Bioscience Database Center, NBDC)のポータル(図1-1-15, URL1-1-28)にまとめられている。 この中のゲノム解析ツールリンク集では、ツールが種類ごとに整理され、簡単 ではあるが日本語の解説が付けられている。また、第1章第4節でも解説する 統合TV(URL1-1-29, 図1-1-16)は、バイオインフォマティクス関連のデ ータベースやツールの使い方を動画で解説していてわかりやすい。例えば、上 記のTaxonomy Browserの使い方に関する動画もある。

バイオインフォマティクス関連のツールやデータベースは変遷の速度が早い ので、最新の状況を知るには書籍よりも、講習会などに参加すると現状が把握 しやすいし、講師に直接質問することで、自分にとって有益なツールやデータ ベースを効率的に知ることができる.定期的な講習会としては、DDBJが開催 するDDBJing (URL1-1-30) やNBDCの統合データベース講習会 (URL1-1-31) がある.また、農業分野としては、育種学会講演会で開催されるバイオイ ンフォマティクス講習や農業生物資源研究所が主催する次世代シークエンサデ ータ処理のためのデータ解析実習があり、果樹研究者にとって非常に有益であ る.農研機構果樹研究所主催の果樹インフォマティクス・キャンプの隔年に1 回程度開催される.



図1-1-15 バイオサイエンスデータベースセンター(NBDC)のトップページ 統合TVやゲノム解析ツールリンク集へのリンクがはられている.



図1-1-16 統合TVのトップページ

引用文献

- Altschul, S. F. *et al.* (1997) Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. Nucleic Acids Research. 25 : 3389-3402.
- Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. Nature Genetics. 25 : 25-29.
- Chagné, D. *et al.* (2014) The Draft Genome Sequence of European Pear (*Pyrus communis* L. 'Bartlett') . PloS one. 9 : e92644.
- Giardine, B. *et al.* (2005) A Galaxy : a platform for interactive large-scale genome analysis. Genome Research. 15 : 1451-1455.
- Huala, E. *et al.* (2001) The arabidopsis information resource (TAIR) : a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Research. 29 : 102-105.
- Jaillon, O. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 449 : 463-467.
- Kaminuma, E. et al. (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. Nucleic Acids Research. 38 : D33-38.
- Kent, W. J. (2002) BLAT-The BLAST-Like Alignment Tool. Genome Research. 12:656-664.
- Neupane, R. *et al.* (2012) Phytozome : a comparative platform for green plant genomics. Nucleic Acids Research. 40 (D1) : D1178-D1186.
- The International Peach Genome Initiative (2013) . The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics. 45 : 487-494.
- Thomas, D. P. *et al.* (2003) PANTHER : A Library of Protein Families and Subfamilies Indexed by Function. Genome Research. 13 : 2129-2141.
- Velasco, R. et al. (2010) The genome of the domesticated apple (Malus × domestica Borkh.) . Nature Genetics. 42 : 833-839.
- Wang, W. J. et al. (2013) The genome of the pear (Pyrus bretschneideri Rehd.) . Genome Research. 23 : 396-408.
- Wu, G. A. et al. (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nature biotechnology. 32 : 656-662.
- Xu, Q et al. (2013) The draft genome of sweet orange (Citrus sinensis) . Nature Genetics. 45: 59-66.