

# DLIB データベースの使用と作成

## 【概要】

- ・Proteome Software 社で公開している DLIB ファイルを利用できます
- ・Scaffold DIA の機能を使うと、Prosit のサイトを経て簡単に DLIB ファイルを作成できます
- ・ Scaffold DIA で DLIB を Library Manager に登録する方法をご紹介

### します

Scaffold DIA では検索の対象とするライブラリとして、Prosit により計算されたピークリスト 並びに保持時間情報を含むデータ(各関連資料にて"DLIB"と記載しています)の使用をお勧め しています。Prosit は Wilhelm と Kuster のグループによって開発されたディープラーニング アルゴリズムです。Scaffold 上で配列データベース(FASTA)を指定する検索と比較すると保持時 間やピーク強度といった内容を持つことにメリットがあり、また DDA ライブラリ検索と比較する と事前の DDA 検索を必要とせず、DDA 結果にとらわれない(DDA で同定できなかったペプチドも ターゲットとできる)点でメリットがあります。ただし DLIB 検索では修飾パターンとしてシス テインのカルバミドメチル化(固定)しか考慮していない事にご注意ください。

本資料では、Scaffold DIAで使用する DLIBの取得方法について以下2つをご案内いたします。

1. Proteome Software 社が公開しているサイトにある DLIB ファイルを利用する

[→P.2]

2. FASTA ファイルを基に Prosit を利用して DLIB ファイルを作成

[→P.4]

また取得した DLIB ファイルについて、

- 3. DLIB 並びに FASTA ファイルを Scaffold DIA にセットする方法
  - [→P.13]
- も関連事項としてご紹介します。



# 1.Proteome Software 社が公開しているサイトにあるファイル を利用するには

Proteome Software 社にて、いくつかの生物種については事前に計算を行った DLIB ファイルと それに対応する FASTA ファイルを公開しています。

<u>https://support.proteomesoftware.com/hc/en-us/articles/360035151172-Prosit-Derived-Sp</u> <u>ectral-Libraries-for-Scaffold-DIA-Searches</u>

#### Library 作成パラメーター

Parameter	Setting
Charge Range	2 - 3
Maximum Missed Cleavages	1
m/z Range	396.4 - 1002.7
Default NCE	33
Default Charge	3

2022 年 12 月末現在、以下の生物種について準備しています(表記は公開サイトの内容に準ずる)

Coronavirus (または plus Human), Arabidopsis thaliana, Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Escherichia coli(strain K-12), Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae

利用したい生物種について、"DLIB"と"FASTA" を両方ともダウンロードしてください。 取得したファイルを Scaffold DIA で利用する方法については、「3. DLIB並びに FASTA を Scaffold DIA にセットする方法(P.13)」をご覧ください。







# 2. FASTA ファイルを基に Prosit を利用して DLIB ファイル作成

Proteome Software 社で予め準備されている DLIB は、対象の生物種が限られています。また FASTA エントリーの対象を Uniprot 内の reviewed (=Swissprot に相当)に限定しています。必要 に応じて公開されているもの以外の生物種を利用したり unreviewed (=TrEMBL に相当)を加えた いなどといった調整が必要な時は、ご自身で DLIB を作成することができます。作成の例として、 生物種 「Oryza sativa」を使ってご説明いたします。

DLIB 作成のステップは以下の通りです。

- 2-A. FASTA ファイルを取得する(例:Uniprot サイト)
- 2-B. Scaffold DIA で Prosit 用 のインプットファイルを作成する
- 2-C. Prosit 公開サイトにてペプチド配列情報から計算を行う
- 2-D. Scaffold DIA で Prosit 出力ファイルと FASTA ファイルから DLIB ファイルを作成する

以降、順にご説明します。

#### 2-A. FASTA ファイルを取得する(例:Uniprot サイト)

FASTA ファイルを準備します。FASTA はどこから取得したファイルでも使用可能ですが、Gene Ontology 情報に紐づけができるデータベースとして、Uniprot データベースの使用をお勧めして おり、本資料でも例として Uniprot サイトからファイルを取得する例をご紹介します。

まず WEB Browser で <u>https://www.uniprot.org/</u> ヘアクセスし、「Proteins」をクリックしま す。





Uniprot に登録されているタンパク質が表示されます。

検索対象を reviewed,しっかりしたアノテーションがついているデータのみにしたい場合は<mark>黄</mark> <mark>色</mark>枠の「SwissProt」あたりをクリックします。Unreviewed も含めた幅広いデータを対象とした い場合は「Uniprot」のあたり(下図<mark>赤枠</mark>)をクリックします。選択しない場合は両方のデータベ ースが選ばれている状態(Uniprot データベース)です。

UniProt BLAST Align Pept	ide search ID mapping	g SF	PARQL UniProtKB	•	Advanced   List
Status Reviewed (Swiss-Prot) (568,363)		B	230,496,50	3 results	l Cuctomizo columno d
Unreviewed (TrEMBL) (229,928,140)	■ Entry ▲	pilos	Entry Name 🔺	Protein Names A	Gene Names 🔺
Popular organisms Human (205.788)	A0A0C5B5G6	8	MOTSC_HUMAN	Mitochondrial-derived peptide MOTS-c[]	MT-RNR1
Rice (149,191)	□ A0A1B0GTW7	8	CIROP_HUMAN	Ciliated left-right organizer metallopeptidase[]	CIROP, LMLN2
Zebrafish (101,302)	A0JNW5	2	UH1BL_HUMAN	UHRF1-binding protein 1- like[]	UHRF1BP1L, KIAA0701, SHIP164
Mouse (86,431) Taxonomy	A0JP26	8	POTB3_HUMAN	POTE ankyrin domain family member B3	POTEB3
Filter by taxonomy	□ A0PK11	Ł	CLRN2_HUMAN	Clarin-2	CLRN2
Proteins with 3D structure (58,487)	A1A4S6	8	RHG10_HUMAN	Rho GTPase-activating protein 10[]	ARHGAP10, GRAF2

Popular organisms			
Human (205,788)			
Rice (149,191)			
A. thaliana (136,466)			
Zebrafish (101,302)			
Mouse (86,431)			
Taxonomy			
алононну			
Filter by taxonomy			

また、左フレームにある「Popular organisms」で表記さ れている生物種に選択したい内容があればクリックする と、該当生物種に限定されたタンパク質のみが表示されま す。目的の生物種がリストにない場合、Filter by taxonomy をクリックします(左図<mark>緑色</mark>の枠)。



生物種名を入力すると、候補が現れます。適切な内容を選択して Search ボタンを押すことで生物種絞り込みが行われます。

Advanced Search		×	А
Searching in			
UniProtKB		▼	
	Taxonomy [OC]		
▼ Taxonomy [OC] ▼	Oryza sativa 🗙	Remove	sto
	Oryza sativa subsp. japonica (Japanese rice/O. sativa/Japonica rice/Rice) [3994		Na
Add Field	Oryza sativa (Rice) [4530]	Search	٩R
(i) Type * in the search box to search for all v	Oryza sativa endornavirus (OsEV) [362693]		2.1
	Oryza sativa aus subgroup [1736659]		, L
	Oryza sativa aromatic subgroup [1736658]		1B
	Oryza sativa subsp. indica (Rice) [39946]		)70
	Oryza sativa Indica Group x <b>Oryza sativa</b> Japonica Group [1050722]		33
	Orvza sativa, Japonica Group x <b>Orvza sativa</b> Indica Group [1080240]		

生物種別のタンパク質リストに変わります。変更後のエントリー数は左側の"Filter by" で 確認ができます。

	n Pept	ide search ID mappi	ng S	SPARQL UniProt	tKB • (taxonomy_id:4530)	
Status						
Reviewed (Swiss-Prot) (5,120)		UniProtk	(B	195,603 r	esults	
Unreviewed (TrEMBL)		BLAST Align M	ap ID	s 🛨 Download	☆ Add View: Cards ○ Table	e 💿 💆
(190,483)		Entry 🔺		Entry Name 🔺	Protein Names 🔺	Gene
Popular organisms		□ <b>A0A0P0X9Z7</b>	÷	CWZF7_ORYSJ	Cysteine-tryptophan	CWZ
Rice (149,191)					domain-containing zinc finger protein 7[]	Os07 LOC
Taxonomy		□ B2ZX90	<b>S</b>	FAS1_ORYSJ	Chromatin assembly factor 1	FSM
4530	×				subunit FSM[]	LOC
Filter by taxonomy						P069 P069
Proteins with		C7IW64	÷	ROS1A_ORYSJ	Protein ROS1A[]	ROS



内容を確認後、ファイル取得を行います。画面上部の「Download」をクリックし、Format として"FASTA(canonical)"を選択します。Compressed / Uncompressed はどちらでも結構ですが、 圧縮形式が Windows 標準では解凍できない"gz"ですので、それを解凍できる環境でない 方は"Uncompressed"を選択して下さい。

全ての項目を選択後、「Go」ボタンを押すとファイル取得(ダウンロード)が実行されます。

U	UniProtKB 195,603 results					
BI	AST Align	Map IDs	🗄 土 Download	🖮 Add View: Cards⊖ Table	e 💿 💆 Cu	
	Entry 🔺		Entry Name 🔺	Protein Names 🔺	Gene Na	
	A0A0P0X92	27 🧯	CWZF7_ORYSJ	Cysteine-tryptophan domain-containing zinc finger protein 7[]	CWZF7, Os07g06 LOC_Os0	
	B2ZX90		FAS1_ORYSJ	Chromatin assembly factor 1	FSM, Os	

Download		×
O Download selected (0)		
<ul> <li>Download all (195,603)</li> </ul>		
FASTA (canonical)		•
Compressed		
Yes		
O No		
	Generate URL for API Preview 10	Cancel Download

#### [参考] 生物種情報を特定するために NCBI の Taxonomy サイトが利用できます。

上記操作で「Popular organisms」で生物種を指定する際、名称の類似性などからターゲットと する対象を限定するのが難しいケースがあります。Popular organisms の候補選択肢として 現れる文字列の後ろに表示されている数字は NCBI の TaxID です。この数字について、NCBI の Taxonomy サイトにて事前に調べておくと候補の中から自らのターゲットとなる生物種を選び やすくなります。以下その操作についてご紹介します。

NCBIのTaxonomyサイト <u>https://www.ncbi.nlm.nih.gov/taxonomy</u> ヘアクセスします。 画面上部の検索枠で、生物種名による検索を行います。キーワードを入れて「Search」ボタンを押します。



検索結果が表示されます。目的の生物種名のハイパーリンクをクリックします。

NIH Nation	nal Library of Medicine enter for Biotechnology Information		Log	ı in
Taxonomy	Taxonomy		Search Search	Help
Display Settings:  ✓ Summar <u>Oryza sativa</u> (Asian cultivated rice), spe <u>Nucleotide</u> Protein	y cies, monocots	Send to: ◄	Related information Nucleotide Protein Assembly	

遷移画面にて、該当生物種のLineage(系統)が表示されます。選択項目が正しい階層・種・株であるかを確認しながら、ターゲットとする項目のハイパーリンクをクリックします。



遷移画面にて、NCBIのTaxonomy ID をチェックします。以下の例であれば「4530」です。



Entrez PubMed Nucleotide Protein Genome St Search for as complete name V Concerner V Iock	Clear	Taxonomy E	BioCollections
<u>Oryza sativa</u>	Entrez	records	
Taxonomy ID: 4530 (for references in articles please use NCBI:txid4530)	Database name	Subtree links	Direct links
ourrent name	Nucleotide	2,289,219	322,013
<b>Oryza sativa</b> L., 1753	Protein	442,430	<u>62,025</u>
Genbank common name: Asian cultivated rice	Structure	269	<u>70</u>
NCBI BLAST name: monocots	Genome	1	1
Rank: species	Popset	1,227	<u>1,077</u>
Genetic code: <u>Translation table 1 (Standard)</u>	Conserved Domains	12	5
Mitochondrial genetic code: <u>Translation table 1 (Standard)</u>	GEO Datasets	21,846	16,106
Plant Plastid)	PubMed Central	32,836	<u>32,836</u>
Other names:	Gene	95,351	<u>149</u>
_common name(s)	HomoloGene	<u>9,787</u>	<u>9,787</u>
red rice, rice	SRA Experiments	98,656	<u>72,606</u>

# 2-B. Scaffold DIAで Prosit 用 のインプットファイルを作成する

FASTA を取得したら Scaffold DIA を起動し、menuの Tools -> Convert FASTA to Prosit CSV を選択します。

🜉 Scaffold DIA	
File Edit View Experiment Export	Tools Help
Filters	Windowing Scheme Wizard DIA Structure Viewer (mzML only)
Show Hidden	7 🍾 Launch Peptide Browser
Organize	Convert BLIB to ELIB Convert MSP/SPTXT to ELIB Convert Prosit/Spectronaut CSV to DLIB Convert TraML to ELIB
Samples	Convert ELIB to BLIB Convert ELIB to OpenSWATH TSV Convert FASTA to Prosit CSV
	Combine ELIB libraries

FASTA を選択し、パラメーターはこだわりがない限り基本的にはデフォルト設定で「OK」 ボタンを押してください。FASTA ファイルと同じ場所に CSV ファイルが作成されます。



Convert FASTA to Prosit CSV			×
Parameters: "cite" This function will <i>in silico</i> d create an input file for Pros <u>Searle et al. 2020</u> .	ligest peptides from your FAST it. If you use this feature, plea	A and se cite	
FAST, cuniprot-organism_Oryza+sati	va+[4530]fasta	Edit	
Charge range:	2 🗘 to 3 🗘		
Maximum Missed Cleavage:		1	$\diamond$
m/z range: 396	5.4 🗘 to 1,002.7 🗘		
Default NCE:		33	$\diamond$
Default Charge:		3	$\diamond$
OK	Cancel		

\*ダイアログ内にも記載がありますように、この機能を利用され論文公開・学会発表をされた 場合、リンクの貼られている論文の引用をお願いしています。

## 2-C. Prosit 公開サイトにてペプチド配列情報から計算を行う

Prositのサイトヘアクセスし、「SPECTRAL LIBRARY」をクリックします(赤枠)。 <u>https://www.proteomicsdb.org/prosit/</u>

We now offer two new Pro fragment intensities predia prediction ( <b>Prosit_TMT_in</b> <b>HCD</b> fragmentation metho We assume all the sequen modification explicitly in ye	sit TMT models that will soon be tion ( <b>Prosit_TMT_intensity_2</b> ) rt_2021). The intensity model v ds but you need to add fragmer ces are fully labeled and you dor pur input files.	e published. One is for <b>D21)</b> and the other is for iRT vorks with both <b>CID</b> and itation column to the input. v't need to add the tmt
P PREDICT	LIBRARIES FAQ	STATUS 👩 💡
Prosit offers high quality MS2 pr prediction. Prosit is part of the P trained on the project's high qua research, please cite "Gessulat,	edicted spectra for any organism roteomeTools ( <u>www.proteometc</u> ality synthetic dataset. When usi Schmidt et al. 2019" <u>DOI 10.10:</u>	n and protease as well as iRT <u>iols.org/</u> ) project and was ng Prosit is helpful for your 38/s41592-019-0426-7.
CE CALIBRATION	SPECTRAL LIBRARY	RESCORING
This task estimates the optima	l collision energy (CE) based on	a given search result. You



最初に Scaffold DIA で作成した Prosit 用の CSV を指定するため「Next」ボタンを押します。

$\sim$	,,	F -F		
$\mathbf{O}$	SV			
O F.	ASTA (comming soon)			
	CSV Format			
	modified_sequence collision	on_energy precurs	or_charge	fragmentati
	M(ox)CSDSDGLAPPQHLIR	15	<sup>2</sup> = 0	н
E	MPQSDPSVEPPLSQETFSDLWK	28	2 F 2	H(
	TCPVQLWVDSTPPPGTR	35	3 ш	ő c
	QSQHM(ox)TEVVR	35	5	0
	<ul> <li>nodified_sequence Use up indicate oxidized Methionin restricted to the range of 7 with carbamidomethylation amino acids.</li> <li>collision_energy Use into precursor_charge Use into fragmentation Either HCC</li> </ul>	oper case letter e with "M(ox)" to 30. Each C n. Prosit does n eger values fro eger values fro 0 or CID, Use u	rs in the colur 2. Sequence le is treated as not support U m 10 and 50 m 1 to 6. upper case let	mn and ength is Cysteine or O as ters.*
	·			

遷移画面にて画面左側のボタンをクリックし、Scaffold DIA で作成した CSV ファイルを指定してから「Next」ボタンを押してください。

•	precursor_charge Use integer values from 1 to 6. fragmentation Either HCD or CID, Use upper case letters.*
*0	nly for TMT model
	uniprot-organism $On/za+sativa+[4530]$ fasta trupsin z3 pco33 $\times$
0	containing peptide sequence, collision energy and precursor charge.
	< BACK NEXT >

# 続いてピーク強度並びに保持時間の予測モデルの選択を行います。適切な内容を選び「Next」 ボタンを押します。各モデルについての詳細は、論文

https://www.nature.com/articles/s41467-021-23713-9

などをご参照ください(概ね名前からご判断いただいて問題ないと思います)。



出力のフォーマットを選びます。ここでは"Generic text"を選択して下さい。選択後"submit"ボタンを押すことで計算が実行されます。

\*フォーマットを.MSP にして、scaffold の変換で .MSP→ELIB を選択した場合も、作成後のファイル は目的の DLIB と同じです。何らかのエラーなどで以下ご案内する方法で DLIB が作成されない場合が ありましたらこちらの迂回策をお試しください。



計算には時間がかかります。Task の ID が表示されますが、この TaskID の情報をどこかに 残しておくと、万が一ブラウザを閉じてしまった場合でも後でアクセスすることが可能です。

例えば以下のように Task ID が示されている場合、



URL は以下の通りとなります(後ろの部分がタスクの ID です)。 https://www.proteomicsdb.org/prosit/task/D907DA69FAC698CA2D23D4BDFBBC5144



計算が終わると以下のようにファイル取得が可能となります。「Download」ボタンを押して ください。ファイルは zip 形式で圧縮されていますので、windows ですと 右クリック→すべて展開、で解凍できます。

	LIBRARIES	FAQ	STATUS		
Prosit offers high quality MS2 predicted spectra for any organism and protease as well as iRT prediction. Prosit is part of the ProteomeTools ( <u>www.proteometools.org/</u> ) project and was trained on the project's high quality synthetic dataset. When using Prosit is helpful for your research, please cite "Gessulat, Schmidt et al. 2019" <u>DOI 10.1038/s41592-019-0426-7</u> .					
Task D907DA69	FAC698CA2D2	23D4BDFE	3BC5144		
Your files are ready.					
DOWNLOAD					

## 2-D. Scaffold DIA で Prosit 出力ファイルと FASTA ファイルから DLIB ファイルを作 成する

Scaffold DIA を起動し、menuのTools -> "Convert Prosit/Spectronaut CSV to DLIB"を選択 します。





時間がかかる可能性がある旨警告するダイアログが現れます。「OK」ボタンを押してください。



Prosit から得た CSV ファイルと、Uniprot サイトから取得した FASTA ファイルを選択して OK ボタンを押します。

Convert Prosit/Spectronaut CSV to Library		
Parameters: Spectronaut CSV/XLS:myPrositLib.csv	Edit	
FASTA:uniprot-organism_Oryza+sativa+[4530]fasta		
OK Cancel		

ファイル作成が開始します。以下ダイアログが消えるとファイル作成が完了しており、Prosit から出力された CSV ファイルと同じ場所に DLIB ファイルが作成されています(拡張子:.dlib)。

🛃 Please wait	×
Reading Spectronaut CSV File	-

# 3. DLIB 並びに FASTA ファイルを Scaffold DIA にセットする 方法

Menu の File -> Open Library manager またはアイコンなどで Library manager を 起動します。

Scaffold DIA	
<u>File Edit View Experiment Export Tools H</u> elp	
🖿 🔄 🔚 🚍 🖶 🖓 🐚 🏦 Summarization:	
Filters Open Library Manager	
Show Hidden	8



"Add Library" ボタンを押します。

Library Manager		×
Name	Status	FASTA
BosTaurus_myPrositLib.dlib	Ready	uniprot-reviewed_yes+AND+orga
whole-cell_IP_lib.blib	Ready	Human_Uniprot_Sprt_Trembl_Isof
BSA_67_myPrositLib.dlib	Ready	uniprot-bsa-filtered-reviewed_yes
arabidopsis_thaliana_prosit_gener	Ready	Arabidopsis thaliana_TAIR10.fasta
Add Library	Remove Library Create L	ibrary Download OK

準備した DLIB ファイルを選択します。続いて、新たに表示された DLIB ファイルの行を選択 した状態で右クリック→ "Associate with FASTA"を選択します。

🧱 Library Manager		×		
Name	Status	FASTA		
OryzaeDB.dlib	Ready			
BosTaurus_myPrositLib.dlib	Ready	uniprot-reviewed_yes+AND+		
whole-cell_IP_lib.blib	Ready	Human_Uniprot_Sprt_Trembl		
BSA_67_myPrositLib.dlib	Ready	uniprot-bsa-filtered-reviewed		
arabidopsis_thaliana_prosit_g	Ready	Arabidopsis thaliana_TAIR10.f		
Add Library Remov	ve Library Create Library	Download OK		

DLIB と対になっている FASTA (あるいは Prosit で DLIB 作成時、最初に取得した FASTA)を 選びます。

🛃 開く							$\times$
Look In:	materialDLIBmaking	$\sim$	t	ŧ	lių.	::	≔
📒 uniprot-o	rganism_Oryza+sativa+[4530]	fasta					
File <u>N</u> ame:	uniprot-organism_Oryza+sat	tiva+[49	530]	fasta			
Files of <u>T</u> ype:	FASTA						$\sim$
			開く			取消	í



#### Library Manager に追加 DLIB がセットされました。

Library Manager		×	
Name	Status	FASTA	
OryzaeDB.dlib	Ready	uniprot-organism_Oryza+sat	
BosTaurus_myPrositLib.dlib	Ready	uniprot-reviewed_yes+AND+	
whole-cell_IP_lib.blib	Ready	Human_Uniprot_Sprt_Trembl	
BSA_67_myPrositLib.dlib	Ready	uniprot-bsa-filtered-reviewed.	
arabidopsis_thaliana_prosit_g	Ready	Arabidopsis thaliana_TAIR10.f	
Add Library Remov	re Library Create Library	Download OK	

検索時にLibrary manager をご利用頂く事で、DLIB と FASTAの選択がスムーズになります。

ご案内内容は以上となります。

● 技術サポート

本資料の内容に関してご質問等ありましたら弊社技術サポートにご連絡ください。

電子メール :support-jp@matrixscience.com

電話:03-5807-7897 ファックス:03-5807-7896