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。

また取得した DLIB ファイルについて、

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。

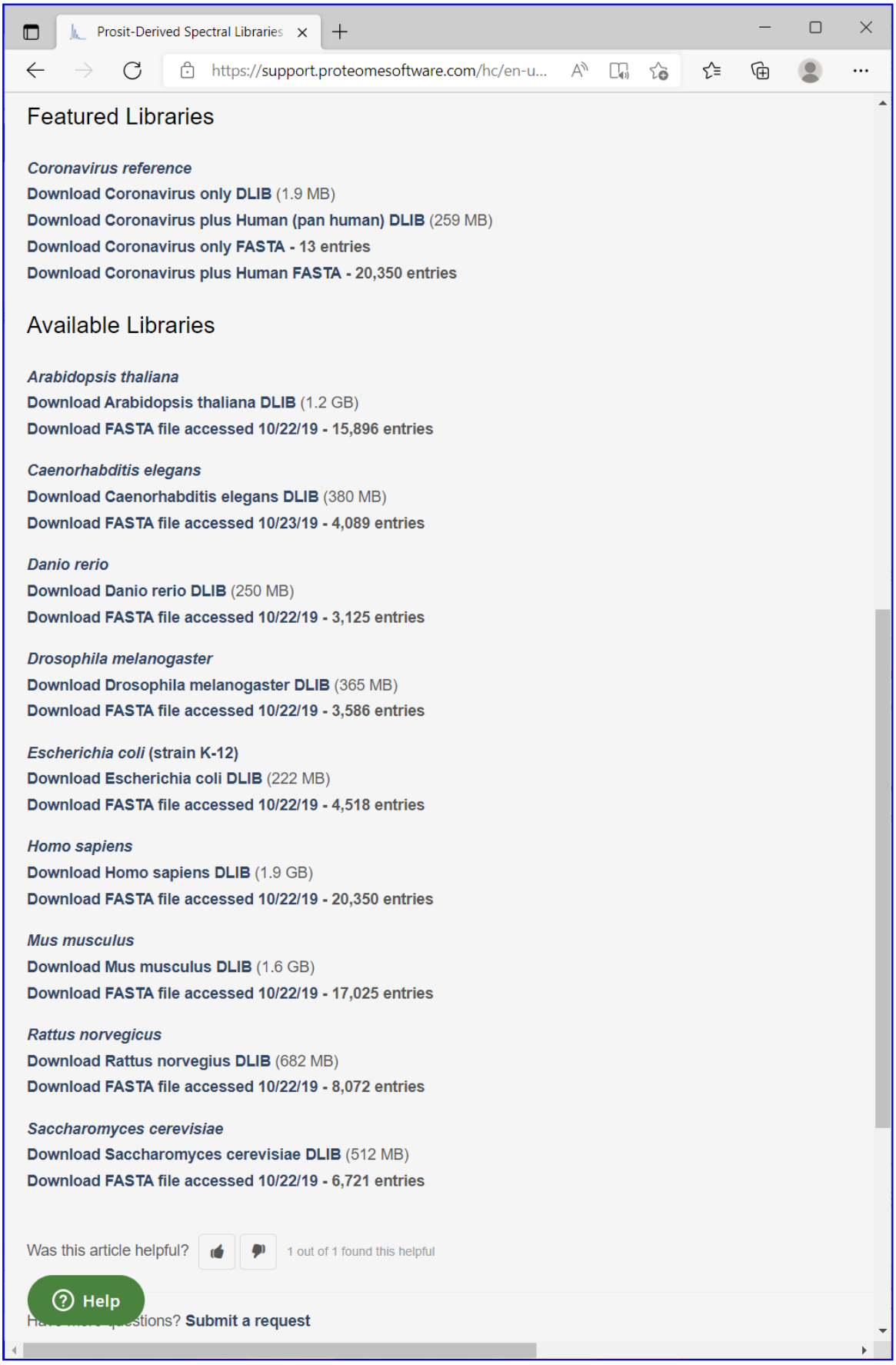
https://support.proteomesoftware.com/hc/en-us/articles/360035151172-Prosit-Derived-Spectral-Libraries-for-Scaffold-DIA-Searches

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。


Prosit-Derived Spectral Libraries
https://support.proteomesoftware.com/hc/en-u...
Featured Libraries
Coronavirus reference
Download Coronavirus only DLIB (1.9 MB)
Download Coronavirus plus Human (pan human) DLIB (259 MB)
Download Coronavirus only FASTA - 13 entries
Download Coronavirus plus Human FASTA - 20,350 entries
Available Libraries
Arabidopsis thaliana
Download Arabidopsis thaliana DLIB (1.2 GB)
Download FASTA file accessed 10/22/19 - 15,896 entries
Caenorhabditis elegans
Download Caenorhabditis elegans DLIB (380 MB)
Download FASTA file accessed 10/23/19 - 4,089 entries
Danio rerio
Download Danio rerio DLIB (250 MB)
Download FASTA file accessed 10/22/19 - 3,125 entries
Drosophila melanogaster
Download Drosophila melanogaster DLIB (365 MB)
Download FASTA file accessed 10/22/19 - 3,586 entries
Escherichia coli (strain K-12)
Download Escherichia coli DLIB (222 MB)
Download FASTA file accessed 10/22/19 - 4,518 entries
Homo sapiens
Download Homo sapiens DLIB (1.9 GB)
Download FASTA file accessed 10/22/19 - 20,350 entries
Mus musculus
Download Mus musculus DLIB (1.6 GB)
Download FASTA file accessed 10/22/19 - 17,025 entries
Rattus norvegicus
Download Rattus norvegius DLIB (682 MB)
Download FASTA file accessed 10/22/19 - 8,072 entries
Saccharomyces cerevisiae
Download Saccharomyces cerevisiae DLIB (512 MB)
Download FASTA file accessed 10/22/19 - 6,721 entries
Was this article helpful?
1 out of 1 found this helpful
Help
Submit a request

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。

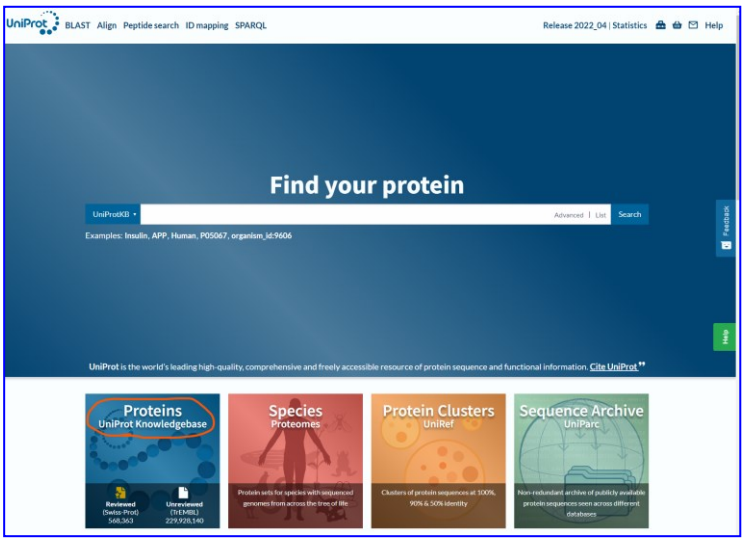
2-A. FASTA ファイルを取得する（例:Uniprot サイト）
2-B. Scaffold DIA で Prosit 用 のインプットファイルを作成する
2-C. Prosit 公開サイトにてペプチド配列情報から計算を行う
2-D. Scaffold DIA で Prosit 出力ファイルと FASTA ファイルから DLIB ファイルを作成する

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）

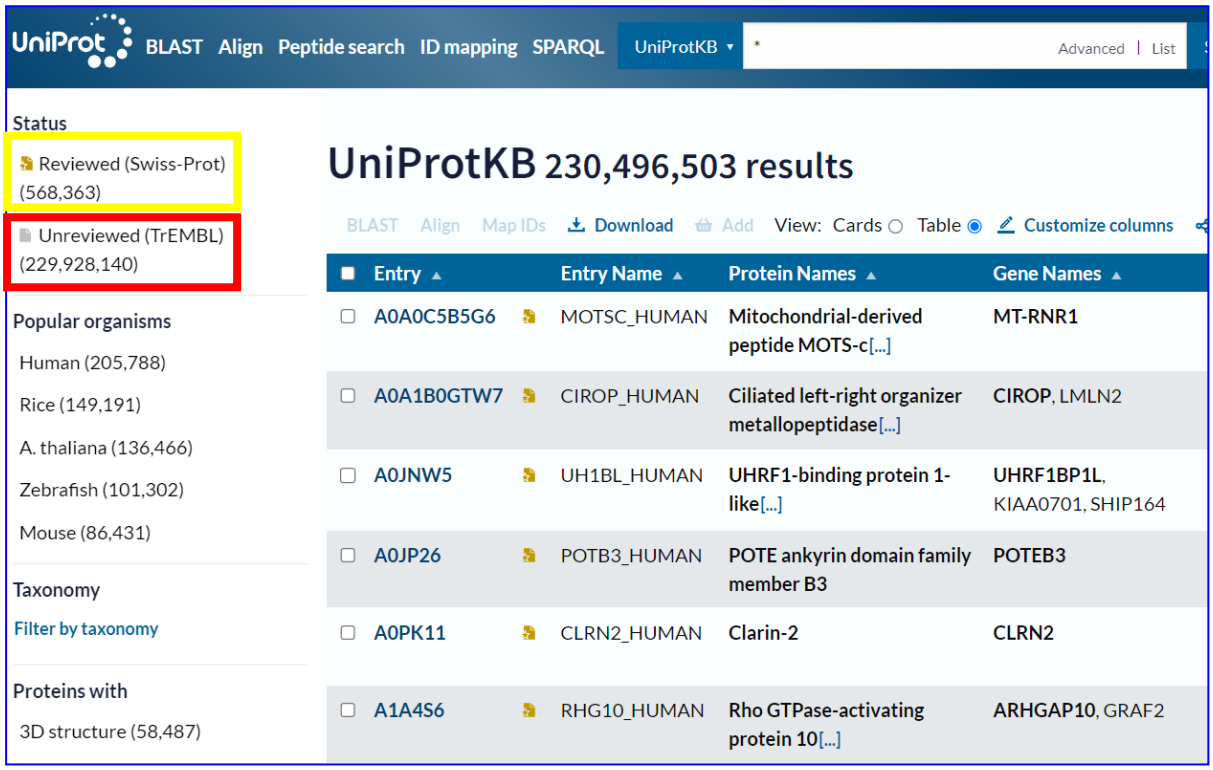
FASTA ファイルを準備します。FASTA はどこから取得したファイルでも使用可能ですが、Gene Ontology 情報に紐づけができるデータベースとして、Uniprot データベースの使用をお勧めしており、本資料でも例として Uniprot サイトからファイルを取得する例をご紹介します。

まず WEB Browser で https://www.uniprot.org/ へアクセスし、「Proteins」をクリックします。



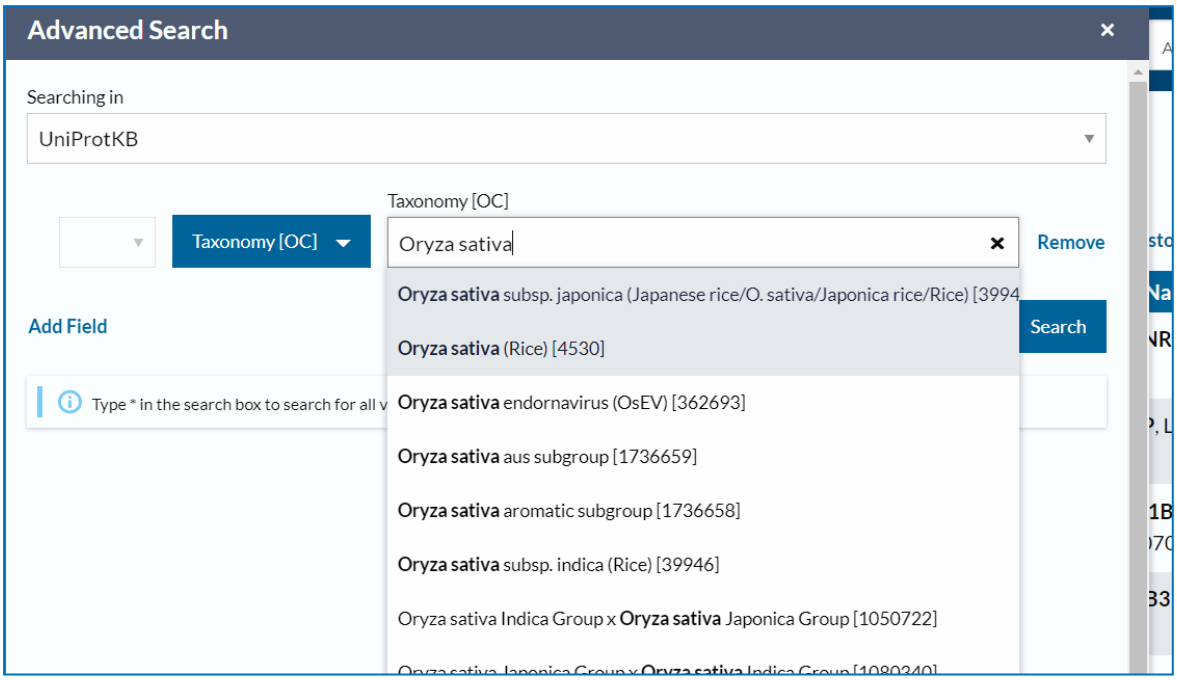
Uniprot に登録されているタンパク質が表示されます。

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黄色枠の「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(下図赤枠)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。

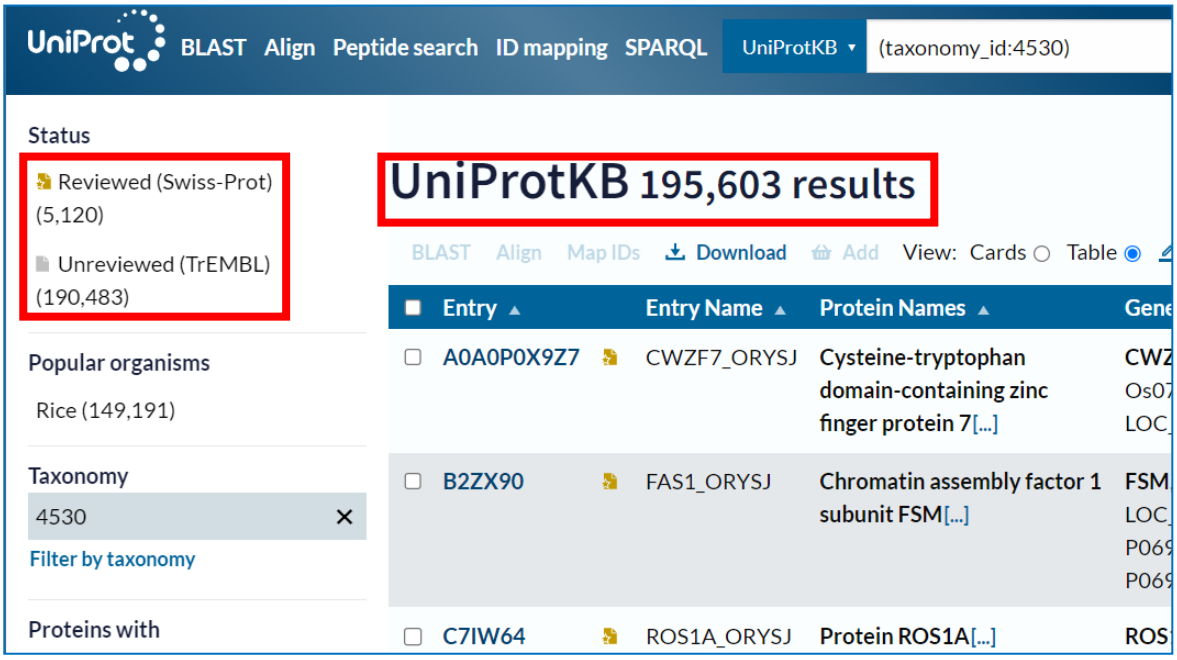


また、左フレームにある「Popular organisms」で表記されている生物種に選択したい内容があればクリックすると、該当生物種に限定されたタンパク質のみが表示されます。目的の生物種がリストにない場合、Filter by taxonomy をクリックします（左図緑色の枠)。

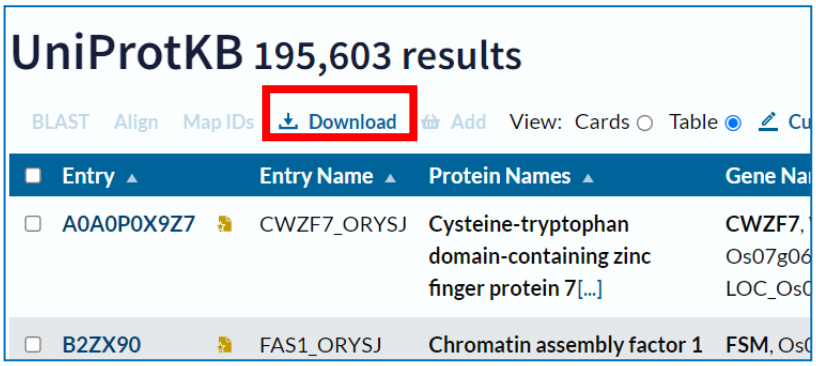
生物種名を入力すると、候補が現れます。適切な内容を選択して Search ボタンを押すことで生物種絞り込みが行われます。

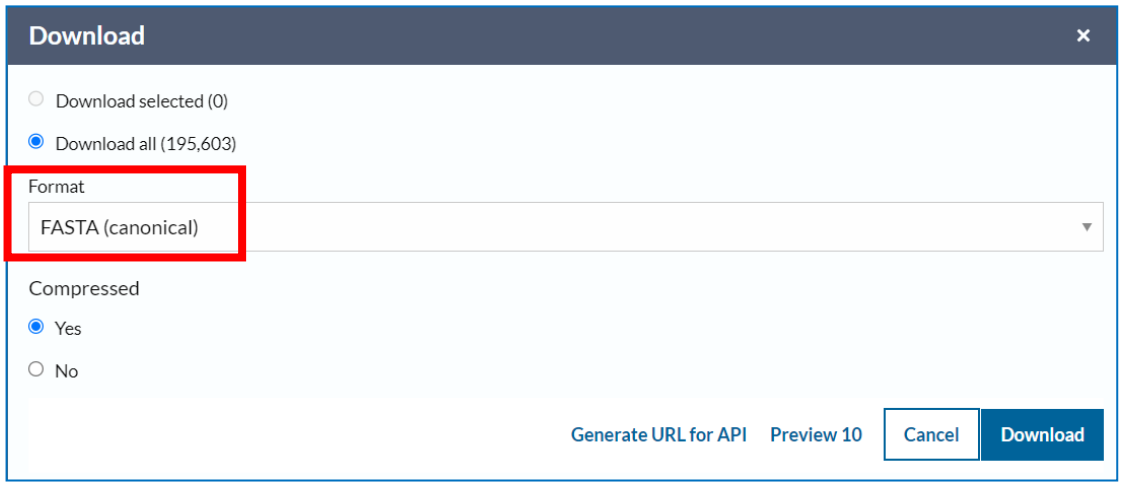


生物種別のタンパク質リストに変わります。変更後のエントリー数は左側の”Filter by”で確認ができます。



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。





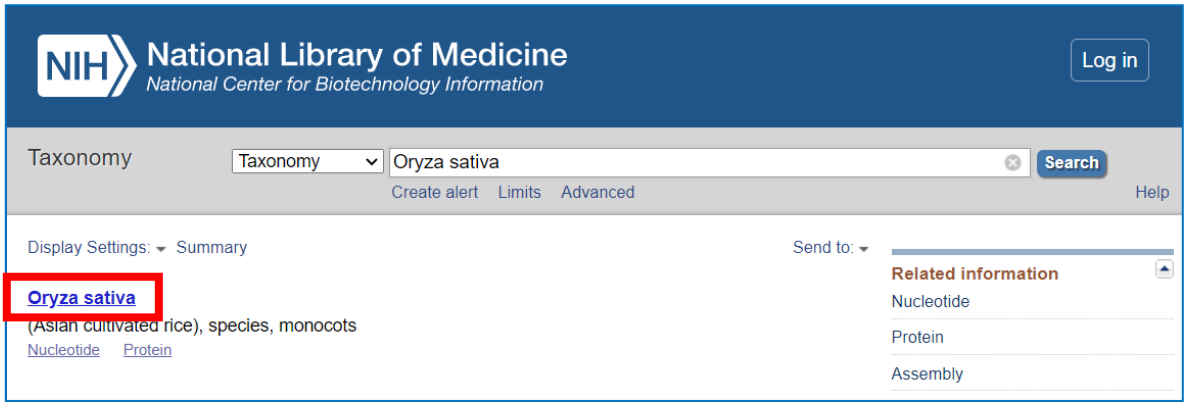
[参考] 生物種情報を特定するために NCBI の Taxonomy サイトが利用できます。

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
https://www.ncbi.nlm.nih.gov/taxonomy
へアクセスします。

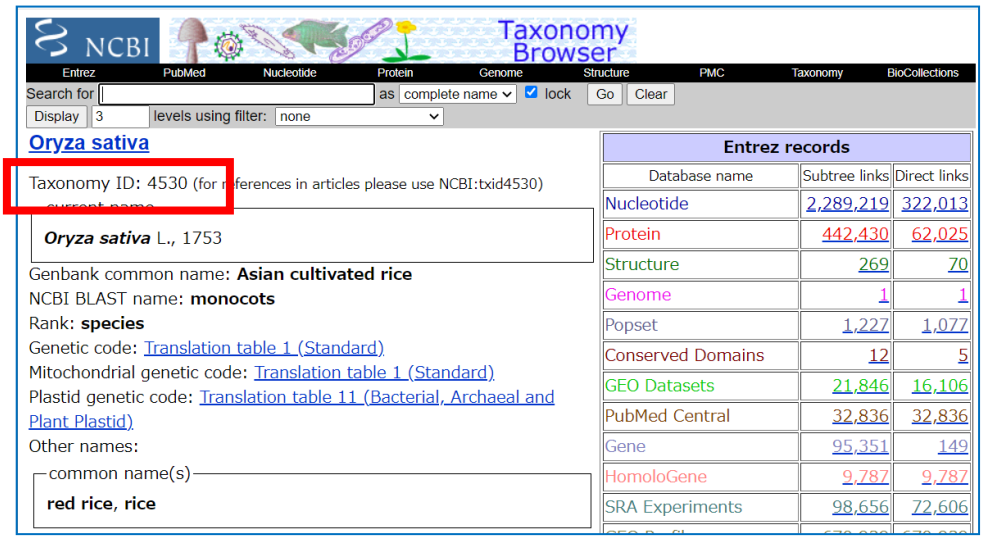
画面上部の検索枠で、生物種名による検索を行います。キーワードを入れて「Search」ボタンを押します。

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。



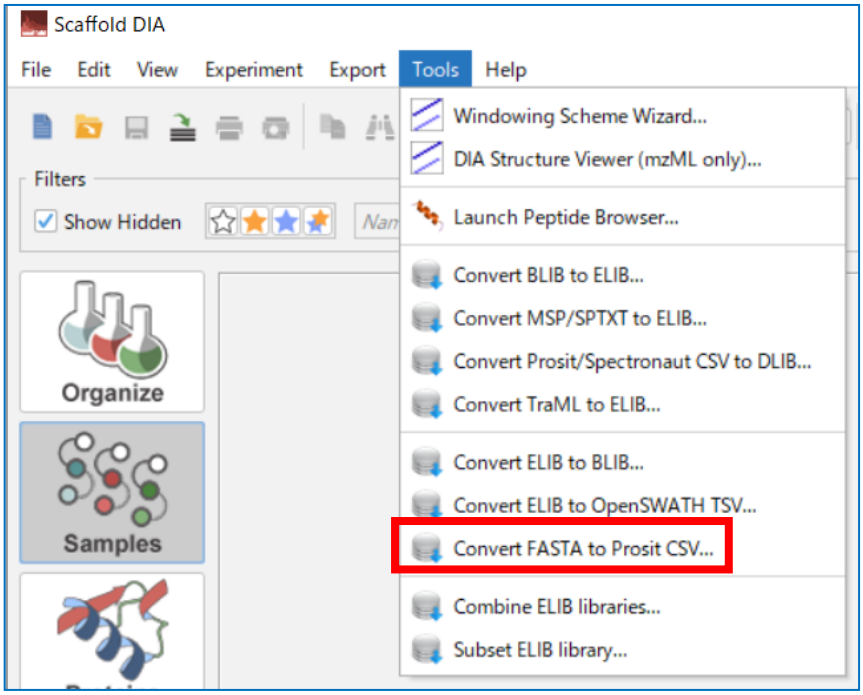
遷移画面にて、該当生物種の Lineage(系統)が表示されます。選択項目が正しい階層・種・株であるかを確認しながら、ターゲットとする項目のハイパーリンクをクリックします。

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。

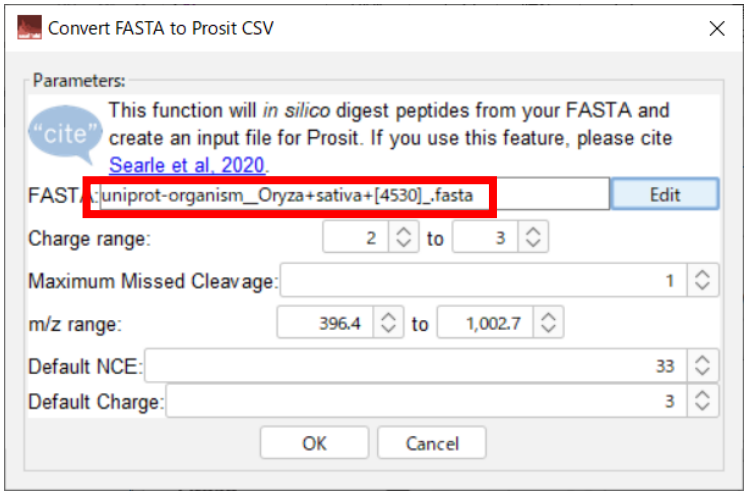


2-B. Scaffold DIA で Prosit 用 のインプットファイルを作成する

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。



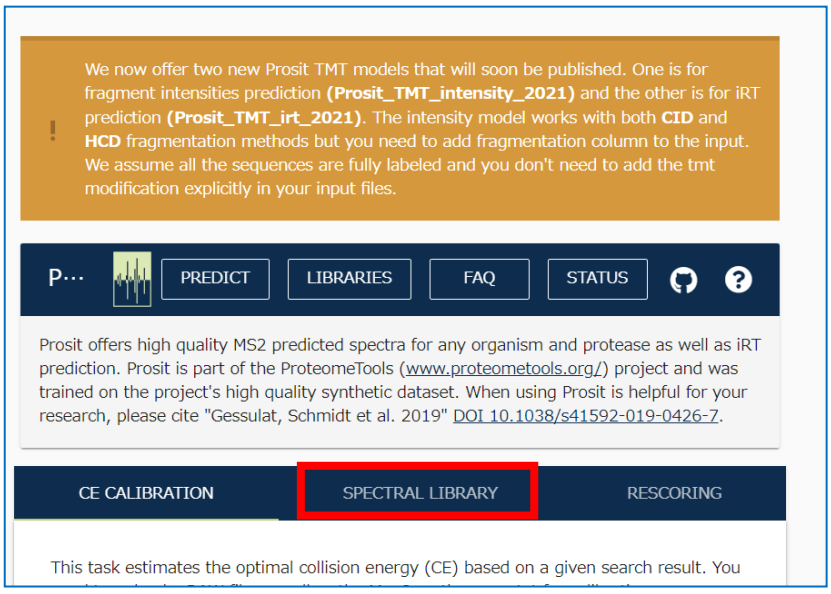
FASTA を選択し、パラメーターはこだわりがない限り基本的にはデフォルト設定で「OK」ボタンを押してください。FASTA ファイルと同じ場所に CSV ファイルが作成されます。



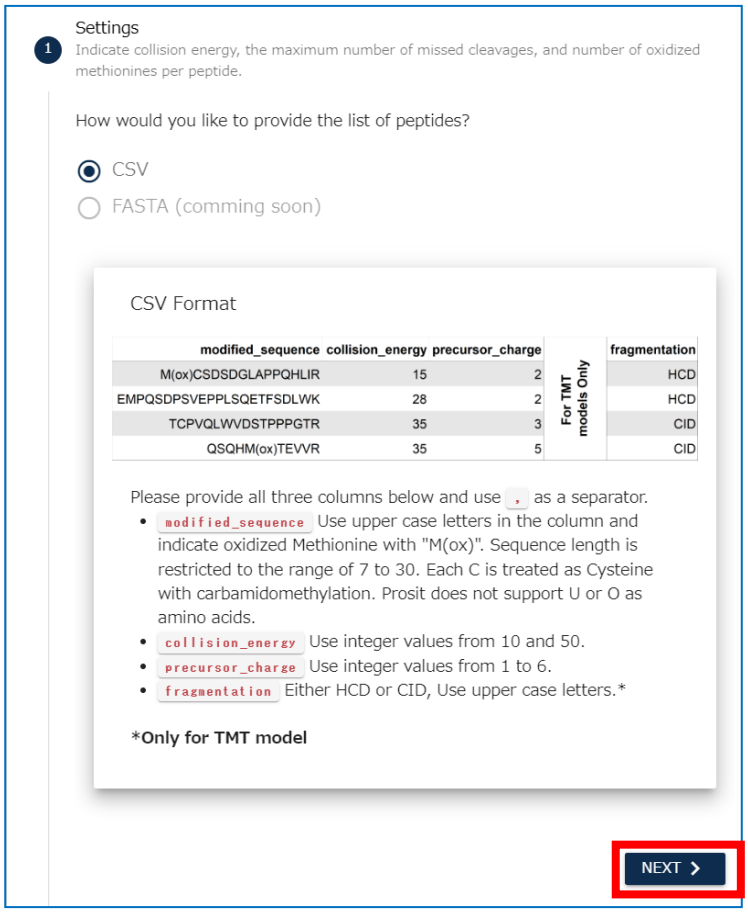
＊ダイアログ内にも記載がありますように、この機能を利用され論文公開・学会発表をされた場合、リンクの貼られている論文の引用をお願いしています。

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

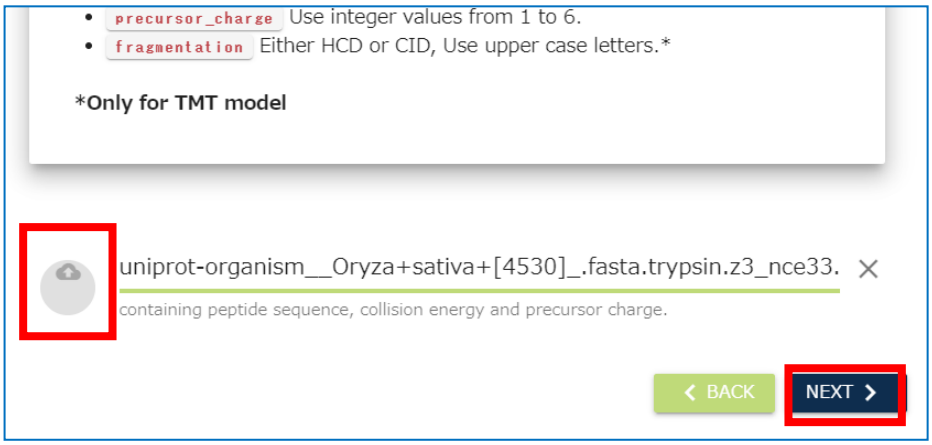
Prosit のサイトへアクセスし、「SPECTRAL LIBRARY」をクリックします（赤枠）。
https://www.proteomicsdb.org/prosit/



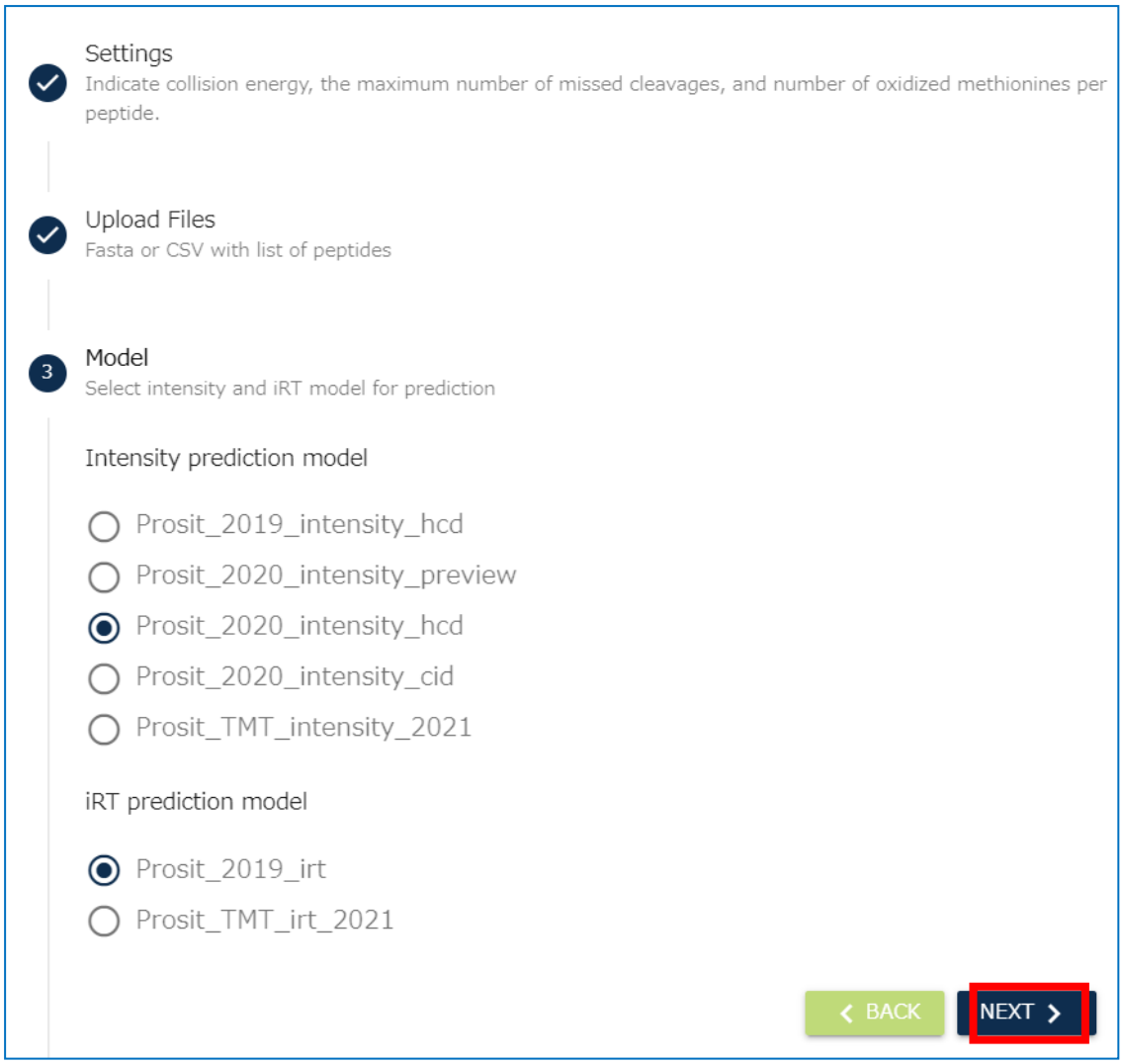
最初に Scaffold DIA で作成した Prosit 用の CSV を指定するため「Next」ボタンを押します。



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。

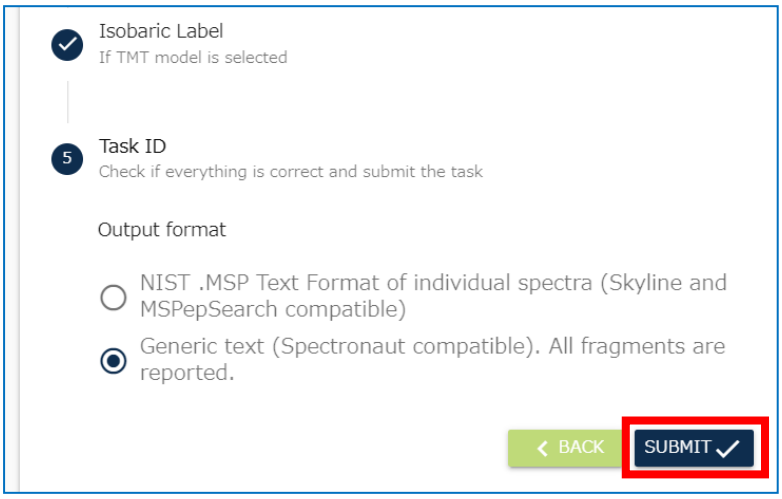


続いてピーク強度並びに保持時間の予測モデルの選択を行います。適切な内容を選び「Next」ボタンを押します。各モデルについての詳細は、論文
https://www.nature.com/articles/s41467-021-23713-9
などをご参照ください（概ね名前からご判断いただいて問題ないと思います）。



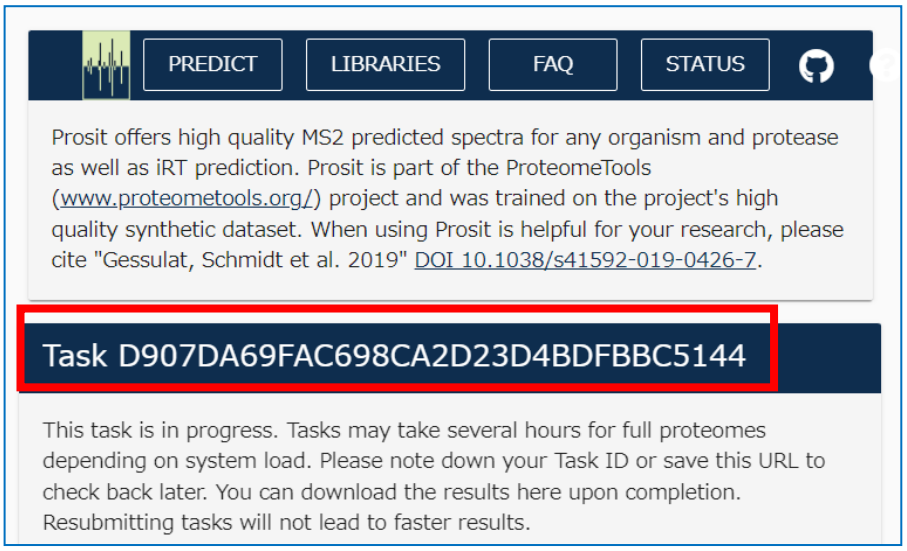
出力のフォーマットを選びます。ここでは”Generic text”を選択して下さい。選択後”submit”ボタンを押すことで計算が実行されます。

*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。

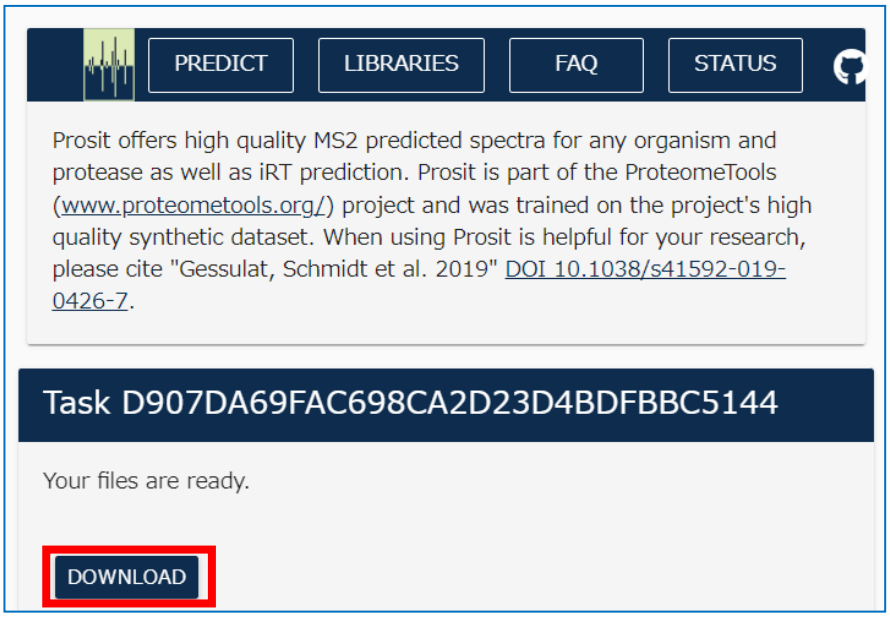
例えば以下のように Task ID が示されている場合、



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）。

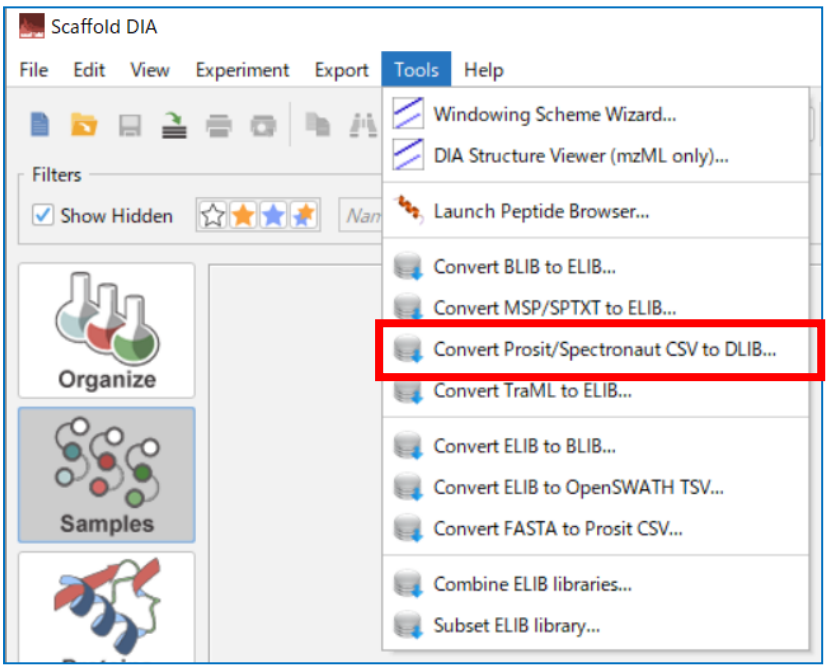
https://www.proteomicsdb.org/prosit/task/D907DA69FAC698CA2D23D4BDFBBC5144

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。

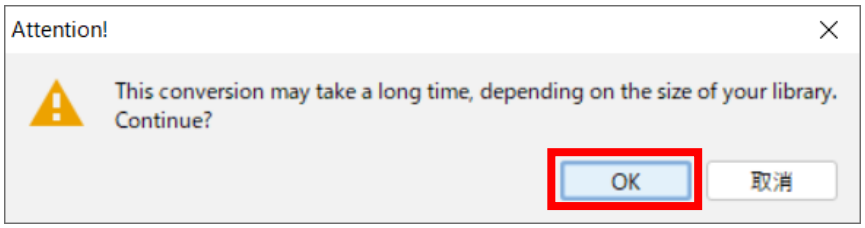


2-D. Scaffold DIA で Prosit 出力ファイルと FASTA ファイルから DLIB ファイルを作成する

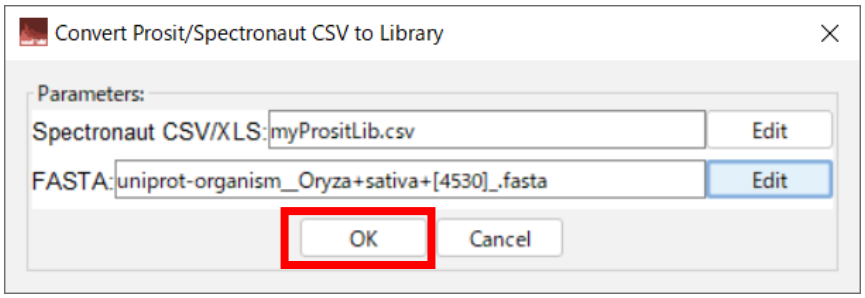
Scaffold DIA を起動し、menu の Tools -> “Convert Prosit/Spectronaut CSV to DLIB”を選択します。



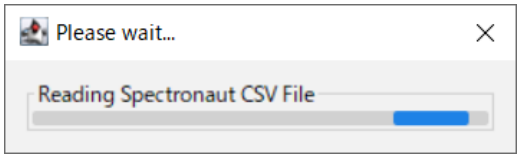
時間がかかる可能性がある旨警告するダイアログが現れます。「OK」ボタンを押してください。



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。

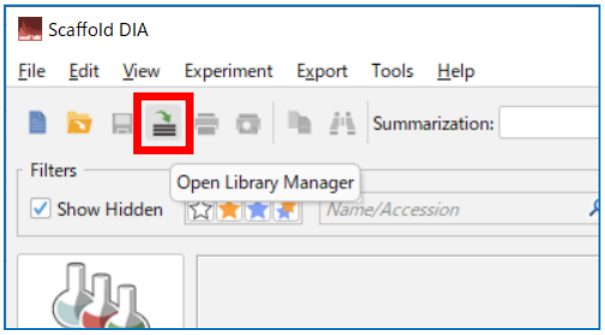


ファイル作成が開始します。以下ダイアログが消えるとファイル作成が完了しており、Prosit から出力された CSV ファイルと同じ場所に DLIB ファイルが作成されています(拡張子 :.dlib)。

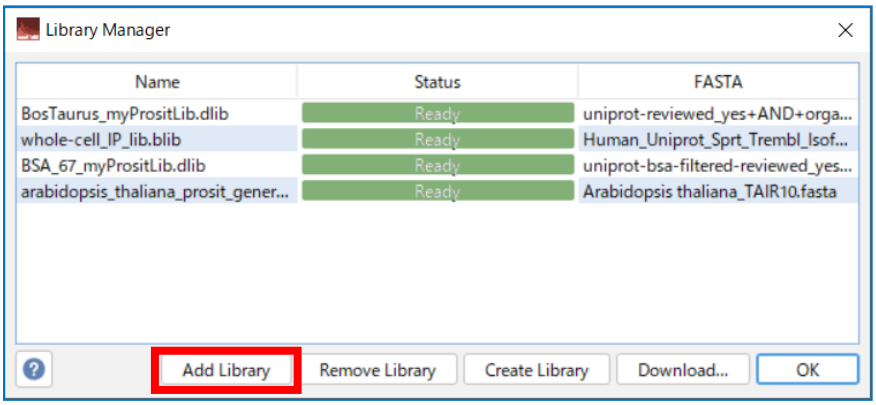


3. DLIB 並びに FASTA ファイルを Scaffold DIA にセットする方法

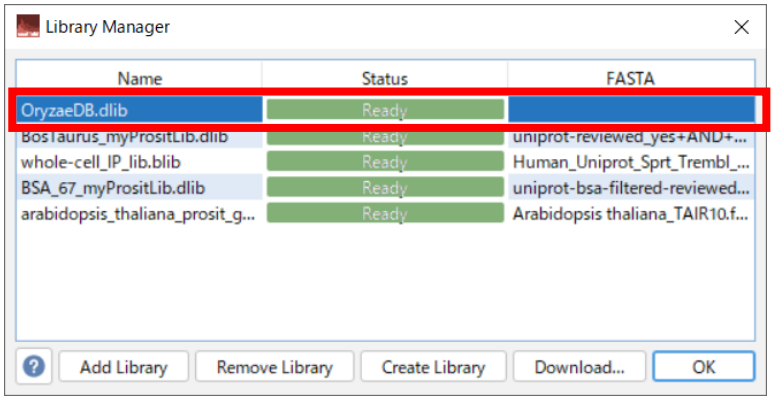
Menu の File -> Open Library manager またはアイコンなどで Library manager を起動します。



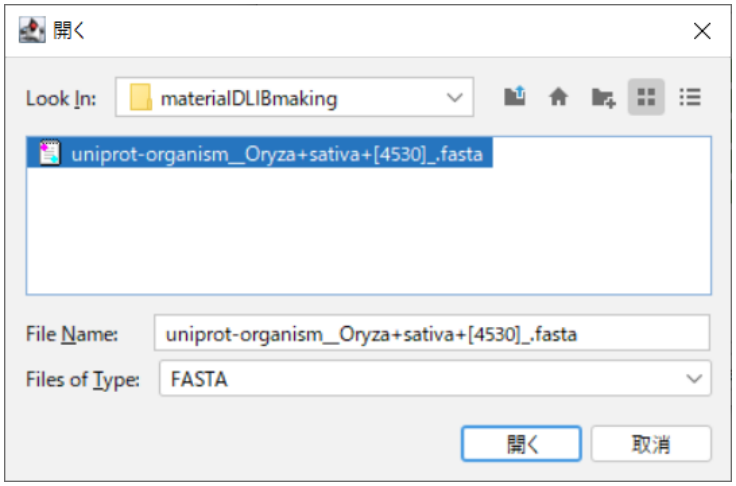
“Add Library” ボタンを押します。



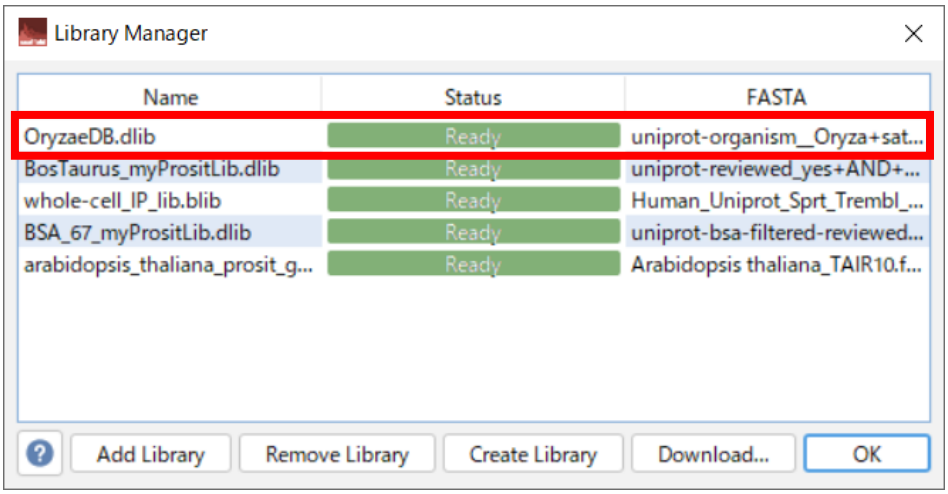
準備した DLIB ファイルを選択します。続いて、新たに表示された DLIB ファイルの行を選択した状態で右クリック→“Associate with FASTA”を選択します。



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。



Library Manager に追加 DLIB がセットされました。



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