検索対象の生物種を予め絞り込んだ UniprotKB データベースの作成手順

【概要】

・NCBIprot の代わりに、生物種別 UniprotKB のご利用をお勧めします。

・本資料では UniprotKB を MASCOT で使用可能にする方法をご紹介します。

MASCOT Server にてよく利用されているデータベース「NCBIprot（旧名称 NCBInr)」は現在ファイルサイズが非常に大きくなりました。そのため主にマシンスペックが不足するコンピュータにおいて NCBIprot データベースを構築できないトラブルが度々発生しています。

根本的にはメモリを中心としたマシンスペックの向上により解決可能ではありますが、NCBIprot のファイルサイズが大きくなるスピードは急速で、ハードウェアが原因の問題が再発する可能性があります。またデータベースの構築ができたとしても、検索に非常に時間がかかってしまいます。

本資料でご説明を予定しているデータベース、「UniprotKB」は２つのデータベースから構成されています。reviewed(内容が精査された)のタンパク質が登録されている SwissProt と、unreviewed(自動アノテーション)の登録が中心の TrEMBL です。

SwissProt は MASCOT にてデータベース名 ”SwissProt”という名称で既に MASCOT にて使用しています。もう一方のデータベース、TrEMBL は NCBIprot と似たコンセプトで集められたタンパク質のデータベースで、知名度も高いことから NCBIprot の代用品として十分ご利用可能です。UniprotKB の特徴として、ファイルを取得する際に生物種をはじめとする様々な絞り込み条件をユーザー側で設定できるメリットがあります。生物種の絞り込みを行う事でファイルサイズが小さくなり、データベース構築時のトラブルを避ける事ができます。

本来であれば、既に MASCOT にあるデータベース「SwissProt」に、「TrEMBL」の内容を加える事で UniprotKB と同一になりますが、毎回複数データベースを組み合わせて検索する操作は若干手間がかかる事から、”UniprotKB”を検索データベースとして利用する事が増えています。UniprotKB をご利用の際には、SwissProt のエントリーも内包している点にご注意ください。

本資料では、UniprotKB で生物種の絞り込みを行ったデータベースを取得し MASCOT にて使用する方法についてご案内しています。UniprotKB を MASCOT で利用するには、

1. MASCOT にて Predefined として定義されているコンテンツを利用して自動的に設定する [→P.2]
2. ご自身で Uniprot サイトからファイルを取得し、データベースをセットする [→P.7]

の 2 通りがあります。

以下、それぞれの方法についてご案内いたします。

1.predefined として定義されているコンテンツを利用して自動的に設定する

2020 年 3 月より、MASCOT 側で Uniprot の定義を predefined としていくつかの生物種で準備しています。
predefined の各データベースの名称は以下のようになっています。

UP5640_H_sapiens

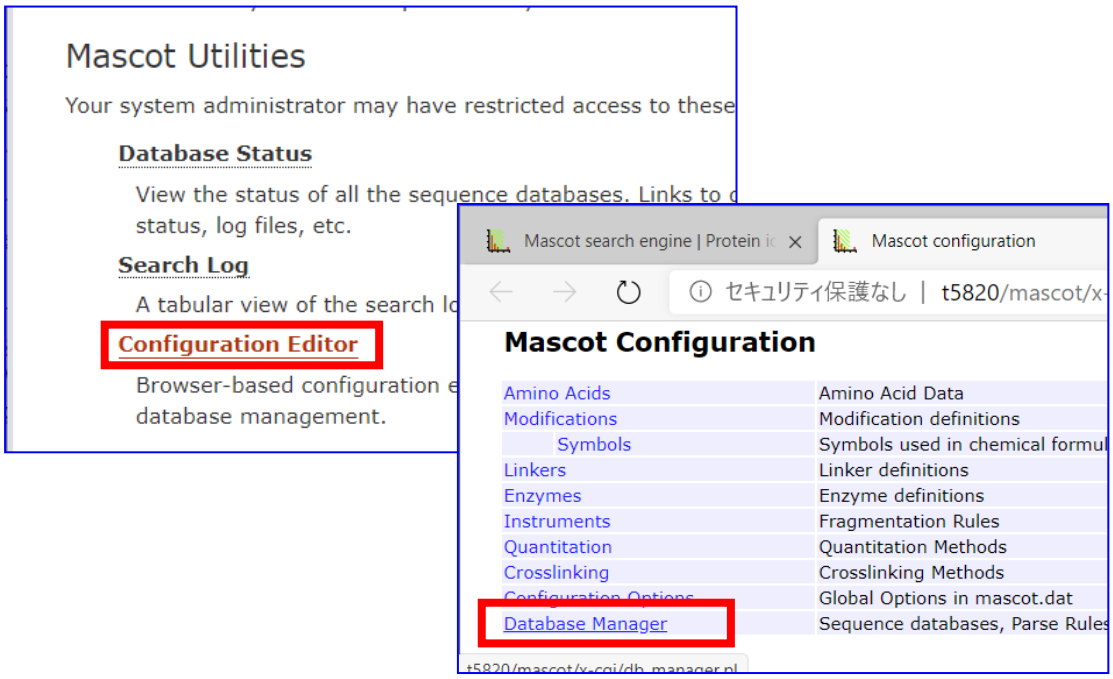
UP は UniProtKB の意味, 5640 は ProteomeID, その後に生物種名という構成です。2020 年 9 月現在、以下 18 種類の生物種が登録されています。

Arabidopsis thaliana , Bos Taurus , Caenorhabditis elegans , Chlamydomonas reinhardtii , Danio rerio , Dictyostelium discoideum , Drosophila melanogaster , Escherichia coli (strain K12) , Homo sapiens , Mus musculus , Mycoplasma pneumoniae , Oryza sativa subsp. japonica , Rattus norvegicus , Saccharomyces cerevisiae , Schizosaccharomyces pombe , Xenopus laevis , Zea mays

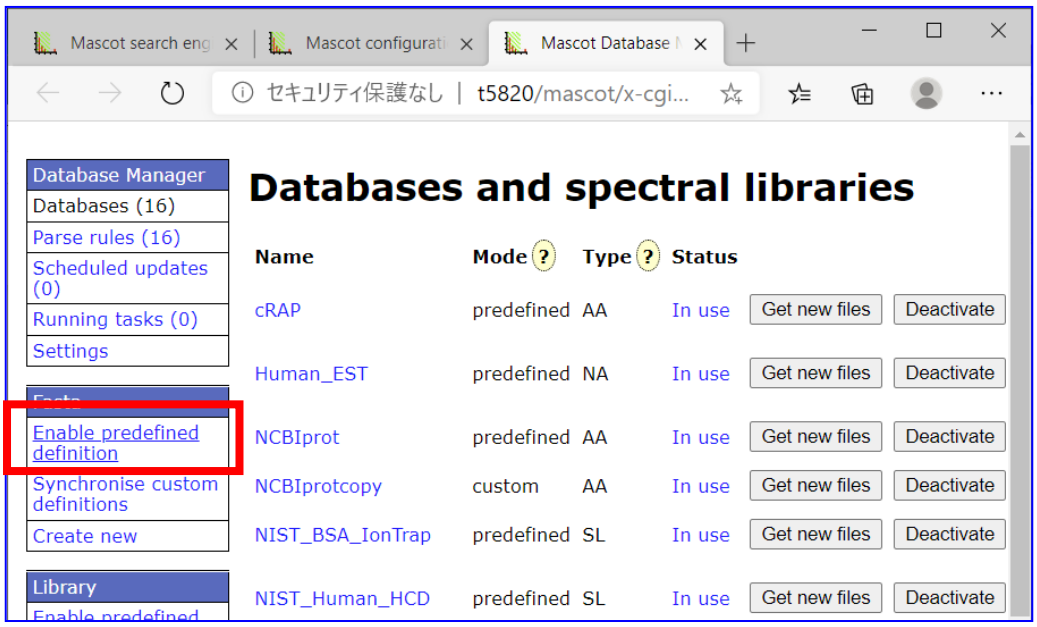
詳細は以下 HP もご参照ください。
http://www.matrixscience.com/search_intro.html#DB

以降、Predefined の内容を MASCOT にセットする操作をご案内いたします。

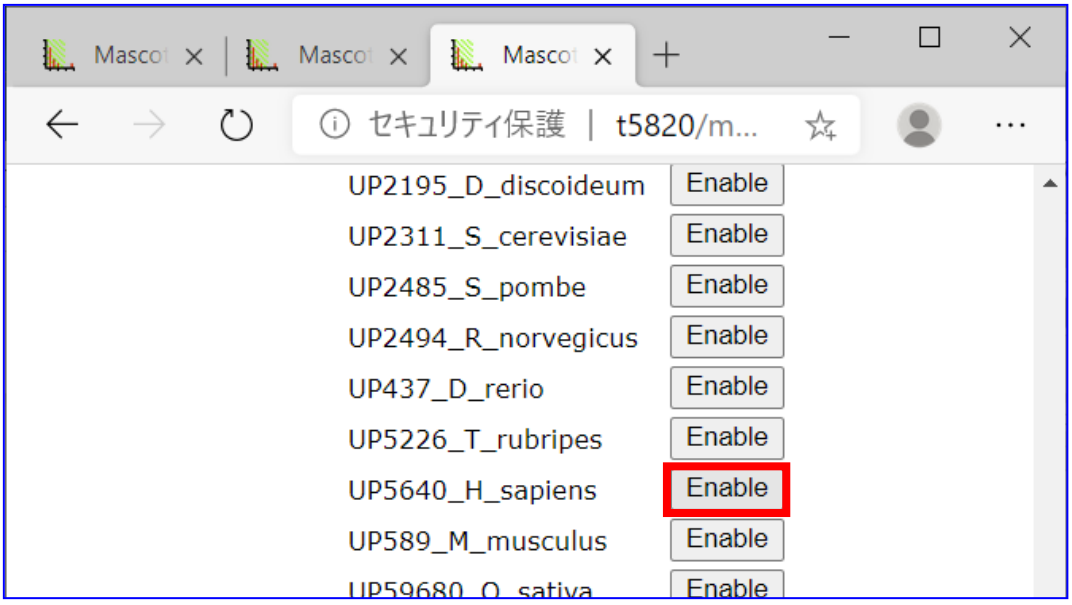
WEB ブラウザで MASCOT の Home 画面を開き、ハイパーリンク ”Configuration Editor”をクリックします。続いて現れる画面で”Database Manager”をクリックします。



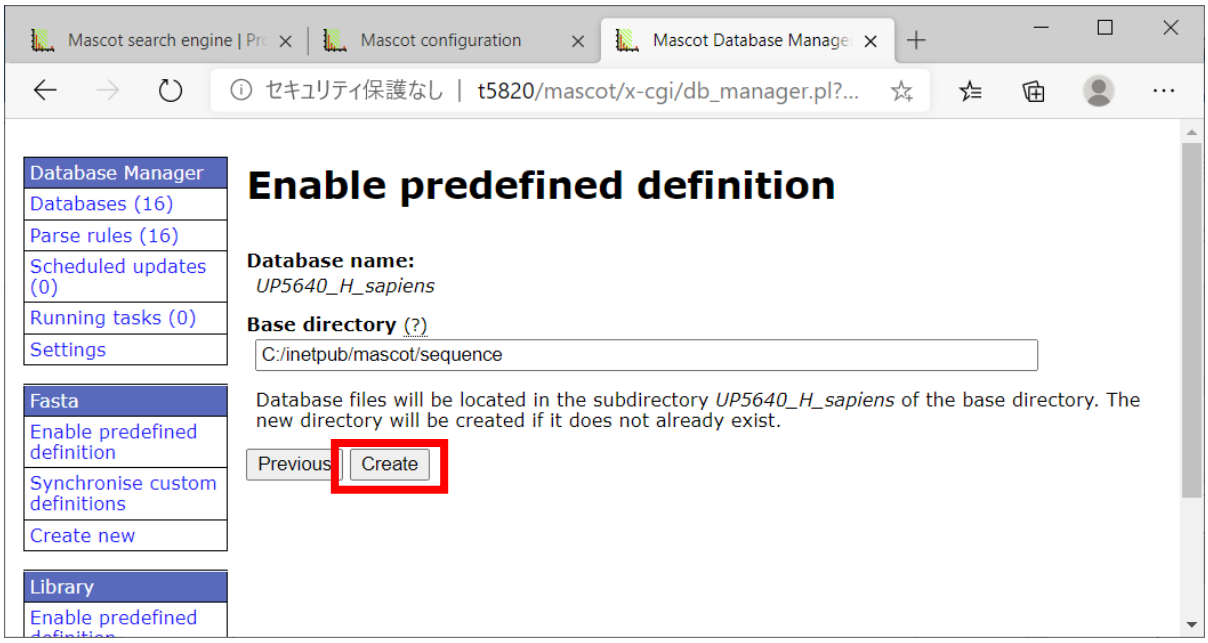
Database manager の設定画面が現れます。
続いて、左フレームにある「Fasta」欄、”Enable predefined definition”をクリックします。



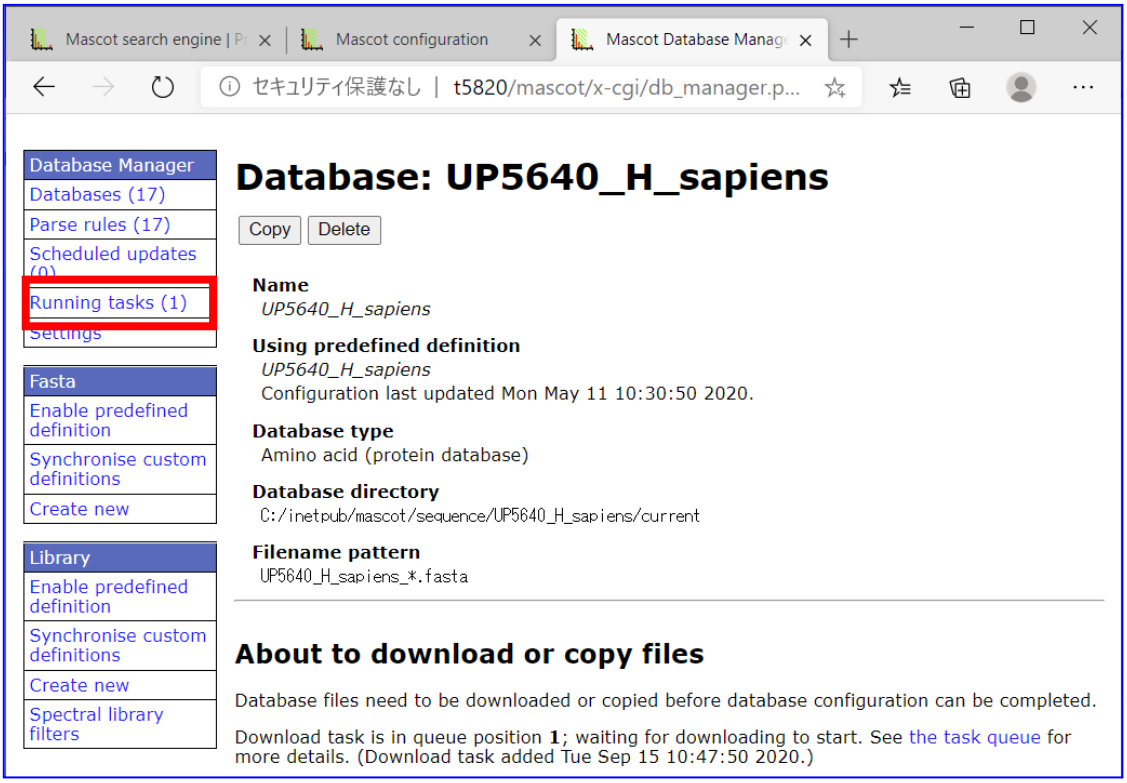
MASCOT 側で準備している、データベース一覧が表示されます。一覧の中で、「UP」から始まるデータベースを探します。自分の目的としている生物種を探し、合うものがあればそのデータベースの行にある”Enable”ボタンを押してください。[例では H_sapiens とします]



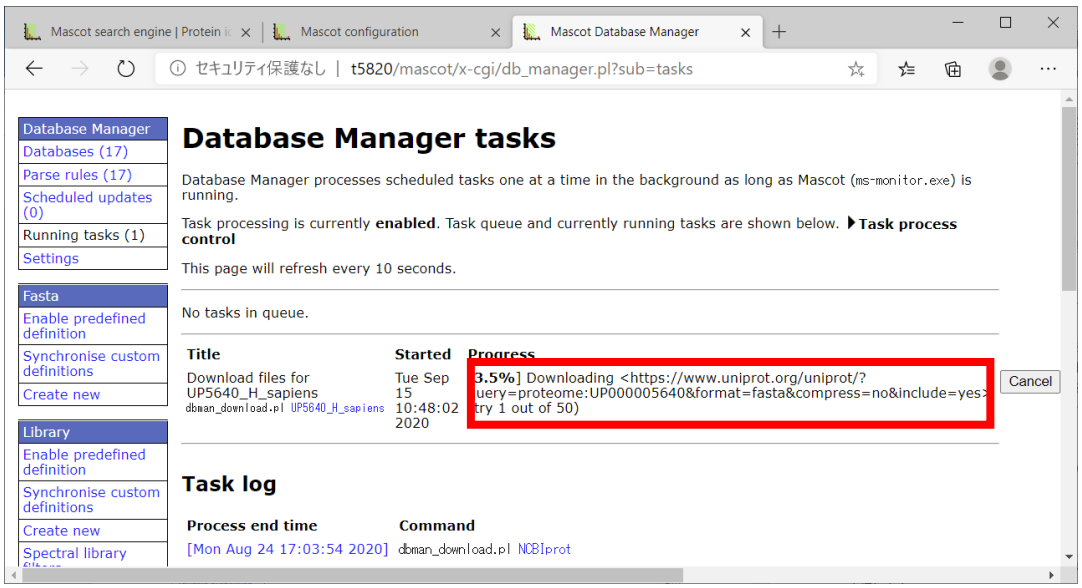
データベースの設置場所について確認する画面が現れます。デフォルト設定で問題がないか、あるいは必要によって設定を変更した上で”Create”ボタンをクリックします。



これで設定は完了です。設定内容の Summary が表示されていますのでご確認ください。また画面一番下に表示されているように、ファイルのダウンロードが開始されます。ダウンロードの進捗をより詳しく確認するためには、”Running tasks”リンクをクリックしてください。

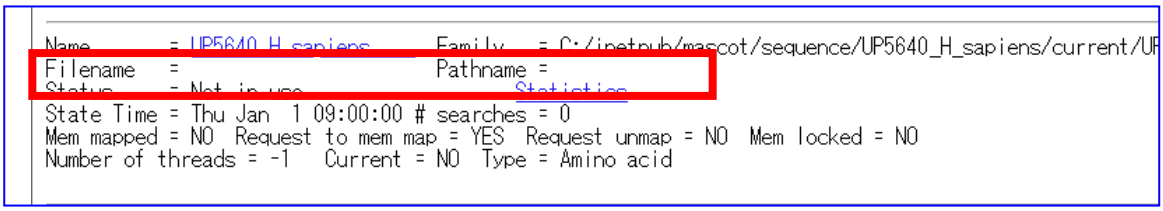


現在ダウンロードされているファイルや既に取得したファイルのサイズなどが表示されます。

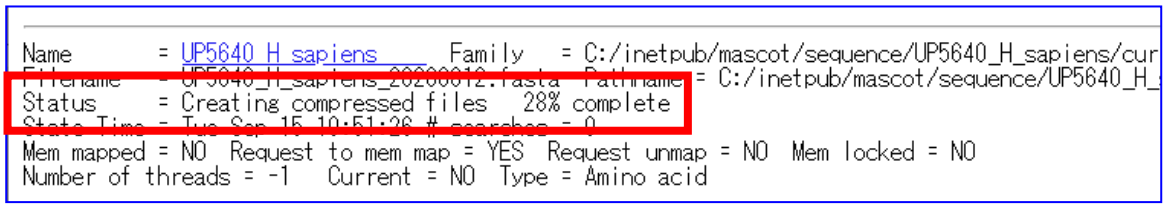


ダウンロードが完了すると、MASCOT にてデータベースの構築が始まります。構築状況は Database Status（MASCOT の Home 画面にリンクがあります）にて確認ができます。データベースの構築を確認する 3 段階の画面についてご案内します。

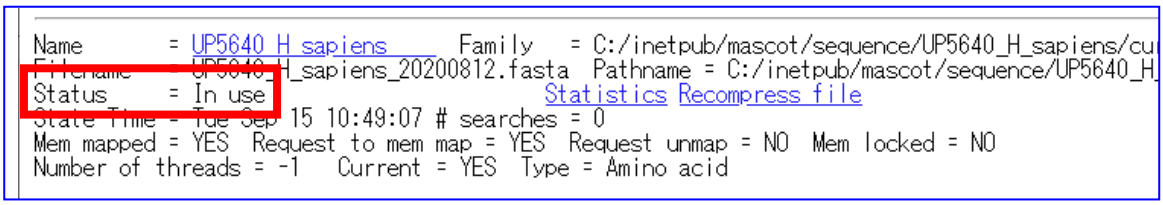
以下のように、”**Filename**”や”**Pathname**”が**空欄**の状態の時は fasta ファイルがセットされていない状態です。ダウンロード後、ファイルを解凍し所定の場所にセットされるまで多少時間がかかる事があります。



構築中の場合、**Status** 項目に”**Creating compressed files N % complete** ” と表示されます。



構築が完了すると、**Status** 項目に”**In use**” と表示されます。In use となれば以降検索する事ができます。



以上で Predefined の定義内容から UniprotKB をセットする操作に関する説明は終了です。

2. ご自身で Uniprot サイトからファイルを取得しデータベースをセットする

予め準備されている設定は生物種が限られています。準備されていた内容以外の生物種、あるいは生物種の組み合わせをご自身でセットするためには、Uniprot サイトにおけるファイルの取得をしたうえで、MASCOT にてセットする必要があります。絞り込みを行う生物種の例として、「Oryza sativa subsp. japonica」を使ってご説明いたします。

始めに、Uniprot サイトよりファイルを取得する方法についてご案内いたします。今回は生物種による絞り込みを行いますが、生物種によっては選択時の候補としてリストアップされる内容からご自身の目的の生物種・階層・株を特定しにくいことがあります。確実に特定するためには NCBI の Taxonomy ID 情報が必要です。そこで、事前に NCBI の Taxonomy ID を確認しておく事をお勧めしています。

これらの手順を、

A. NCBI の Taxonomy ID を確認する（必須ではありませんが、実施をお勧めいたします）
→ P.8

B. ファイルの取得
→ P.10

にてご案内しています。

また、Uniprot から取得したファイルについては簡単に MASCOT にてセットする事ができます。Uniprot のデータベースは使用される頻度も高いことから、MASCOT ではすぐに使用するためのフォーマットを予め準備しており、作成時にはそのフォーマットを選択する形で進行します。作成後、Database Status 画面などで構築状況を確認します。

これらの手順を

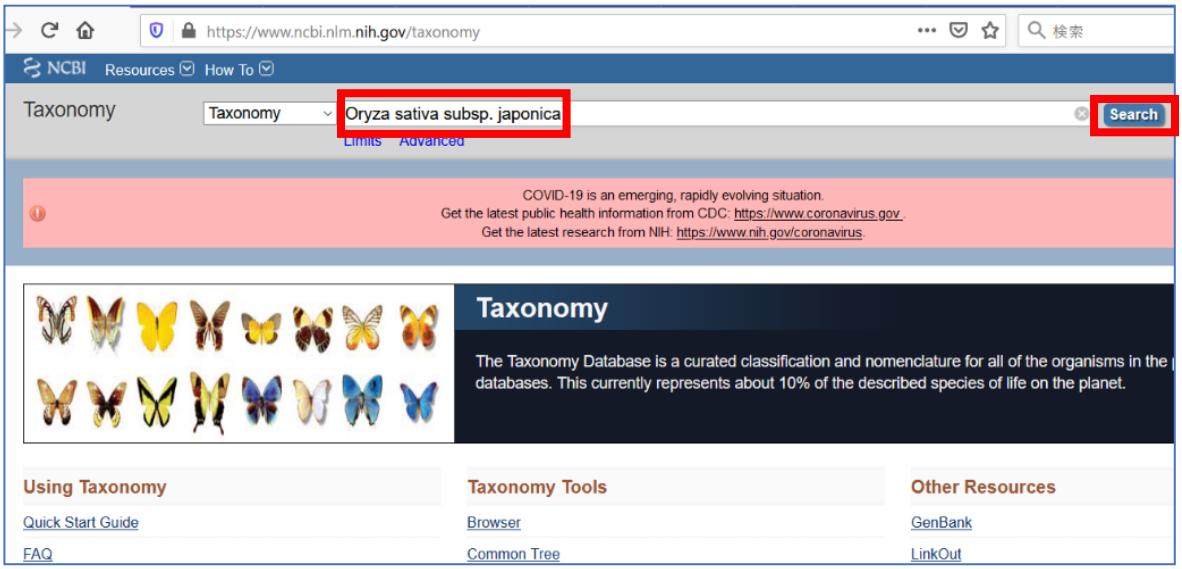
C. MASCOT でのデータベース設定
→ P.13

D. データベースの構築確認
→ P.18

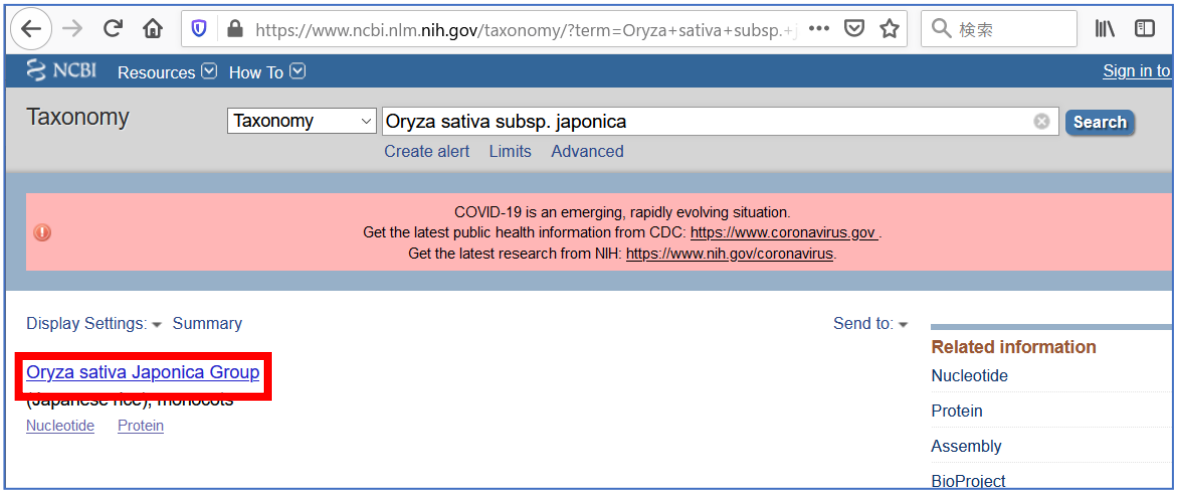
にてご案内しています。

A. NCBI の Taxonomy ID を確認する（必須ではありませんが、実施をお勧めいたします）

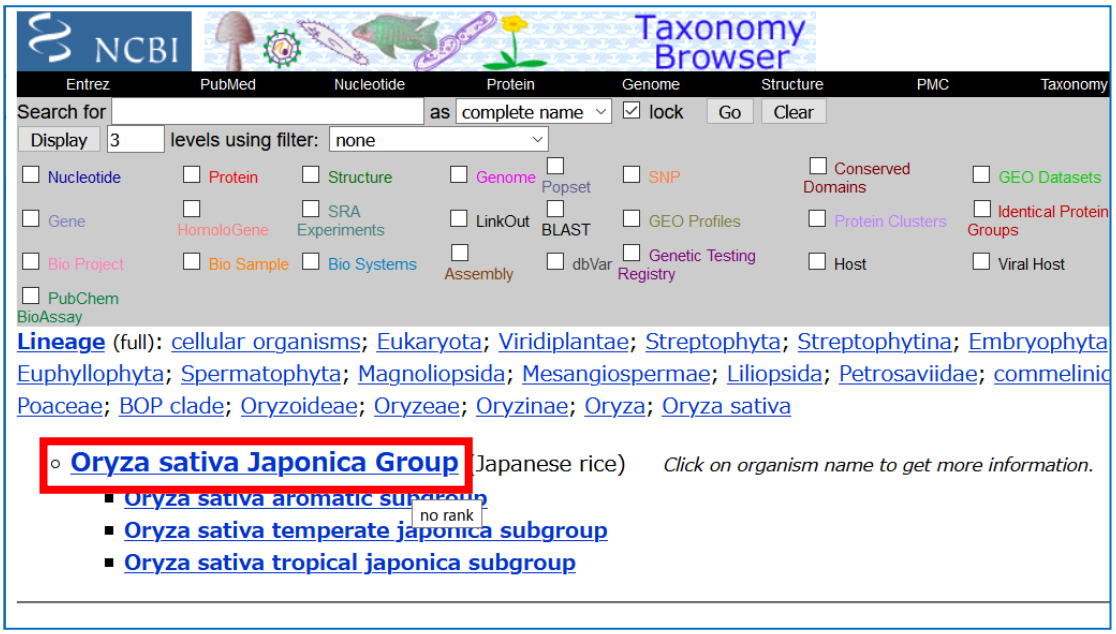
NCBI の Taxonomy サイト
https://www.ncbi.nlm.nih.gov/taxonomy
へアクセスします。
画面上部の検索枠で、生物種名による検索を行います。キーワードを入れて「Search」ボタンを押します。



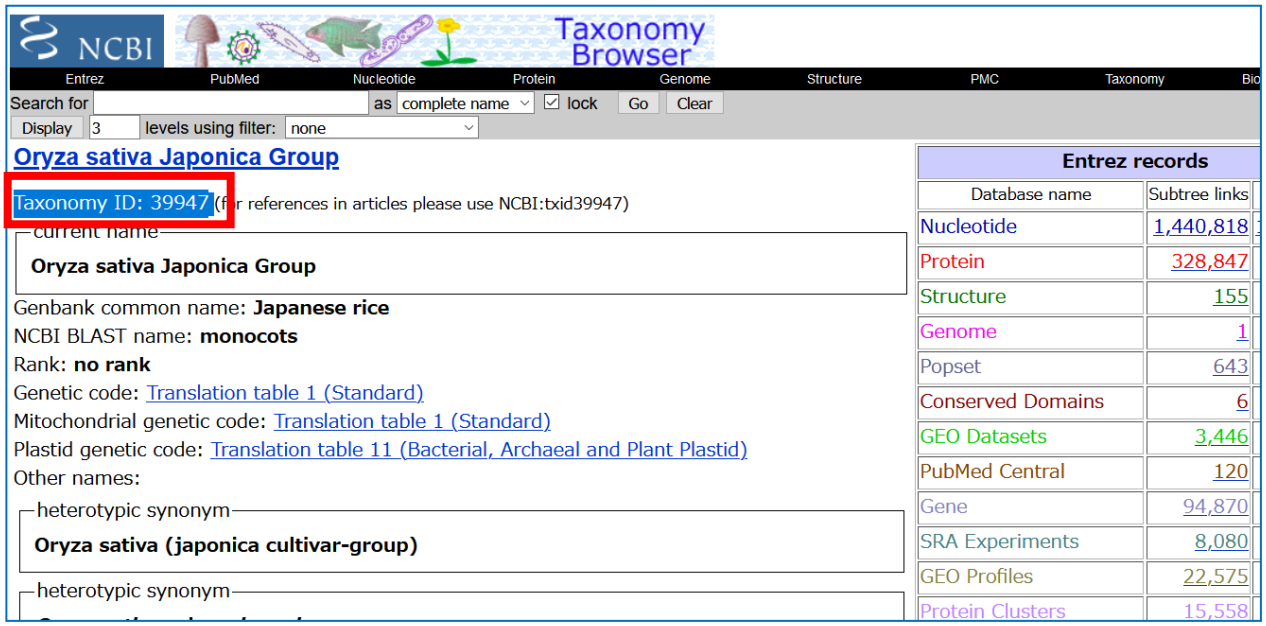
検索結果が表示されます。目的の生物種名のハイパーリンクをクリックします。



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「**39947**」です。



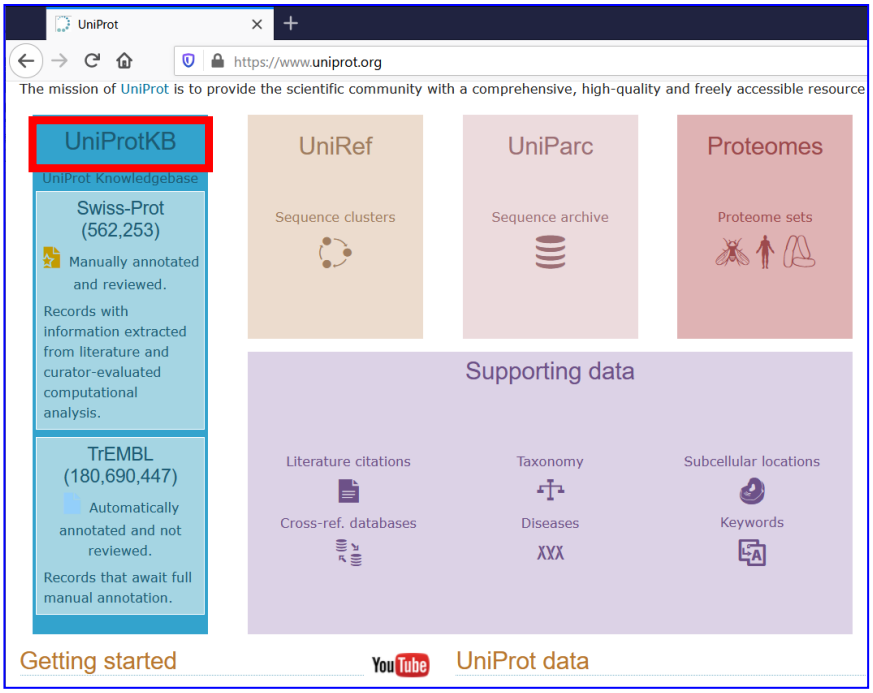
以上の操作で NCBI の Taxonomy ID を確認しておきます。

B. ファイルの取得

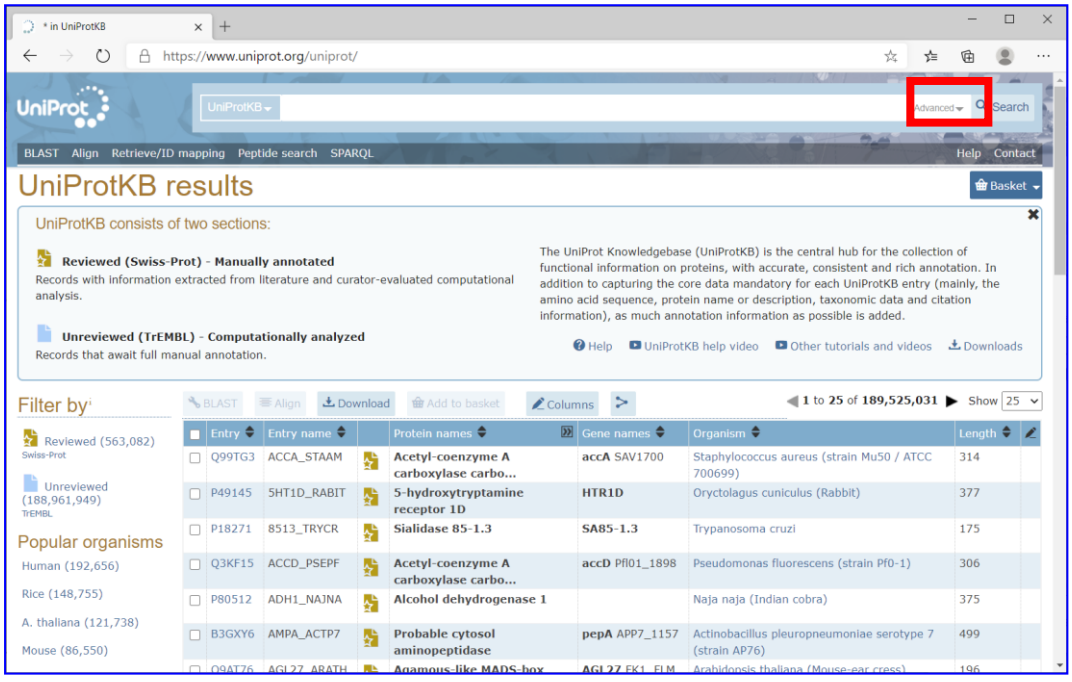
UniprotKB のサイトから、生物種を中心とした条件を絞り込んだ FASTA ファイルを取得します。

まず WEB Browser で https://www.uniprot.org/ へアクセスします。

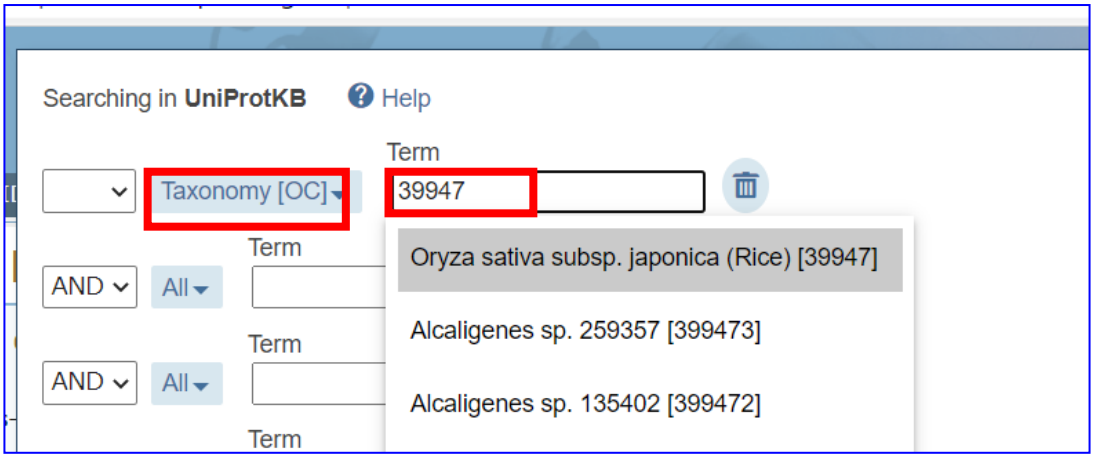
現れた画面の左フレームにある、「Uniprot」のあたり(下図**赤枠**)をクリックします。

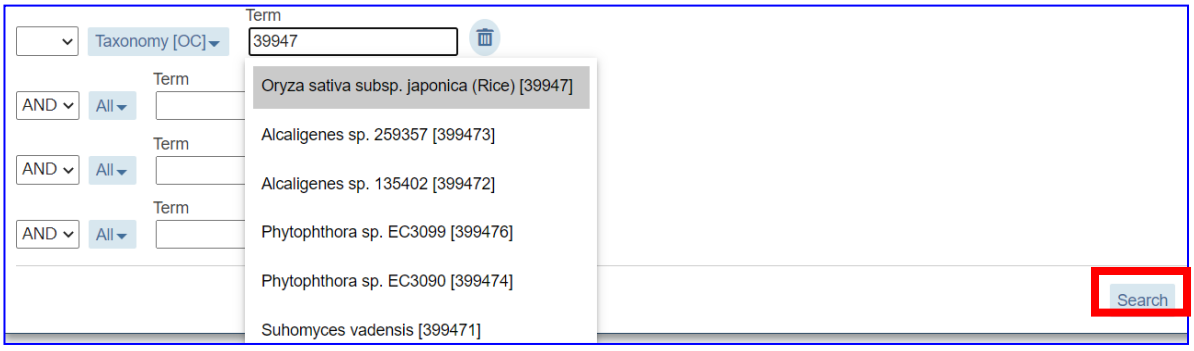


Uniprot のタンパク質一覧が現れます。画面上部にある検索枠の右側、”**Advanced**” をクリックします。



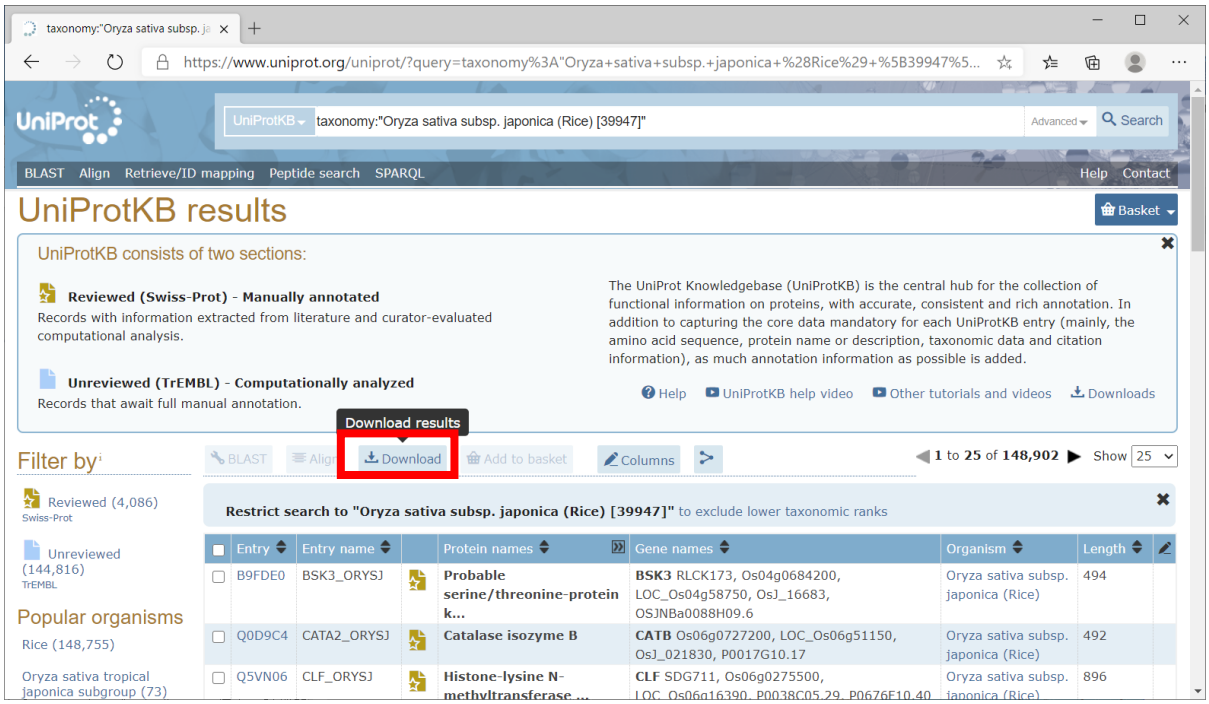
下図、絞り込み条件の項目選択で、**Taxonomy[OC]**を選択します。続けて隣の枠に、生物種名または事前に調べた NCBI の Taxonomy ID を入力します。しばらくすると入力項目に合致する項目が候補として下側に表示されるので、該当項目を選択し、クリックします。リストから項目を選択する際、**名称の後ろの［ ］内の数字が Taxonomy ID である**事に注目してください。類似名称の項目が多数リストアップされた場合、**Taxonomy ID を目安に項目を特定**してください。項目が正しく選ばれていることを確認後、ダイアログ右下の「**Search**」ボタンを押します。





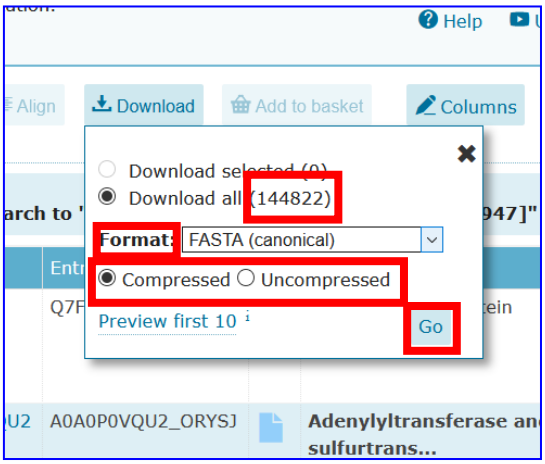
［次頁に続きます］

生物種登録情報が絞り込み条件と合致するタンパク質の一覧が現れます。続いてファイル取得のため、リスト上部にある「**Download**」をクリックします。



ダウンロードに関する設定項目が現れます。

Format ：「FASTA」 を選択し、「**Go**」 ボタンを押すとファイルの取得が始まります。

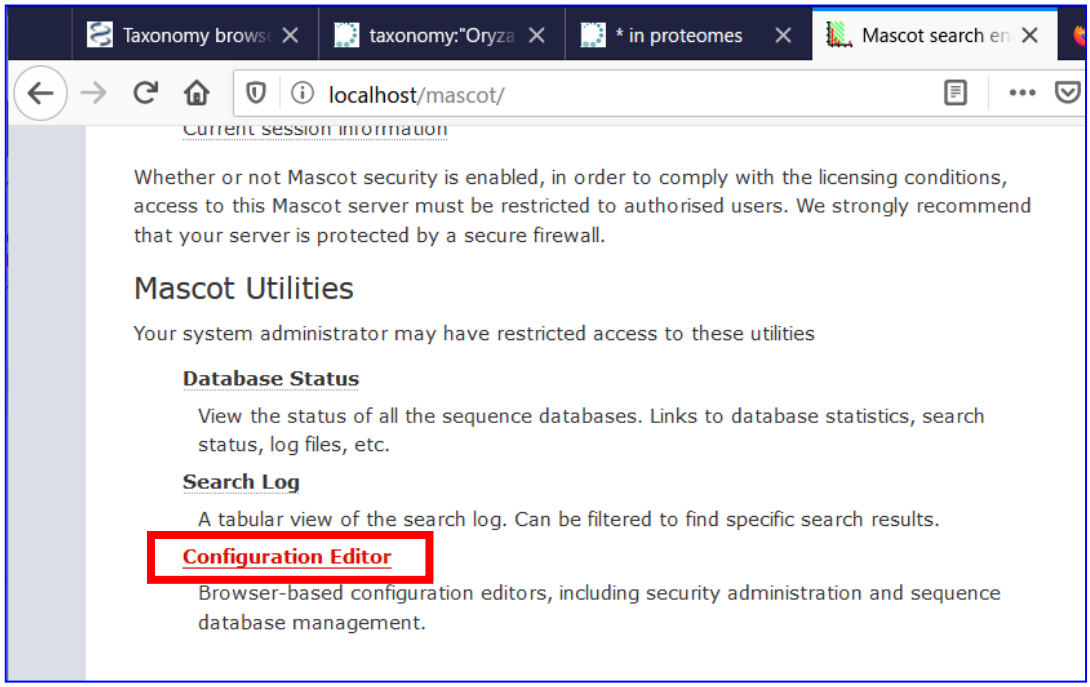


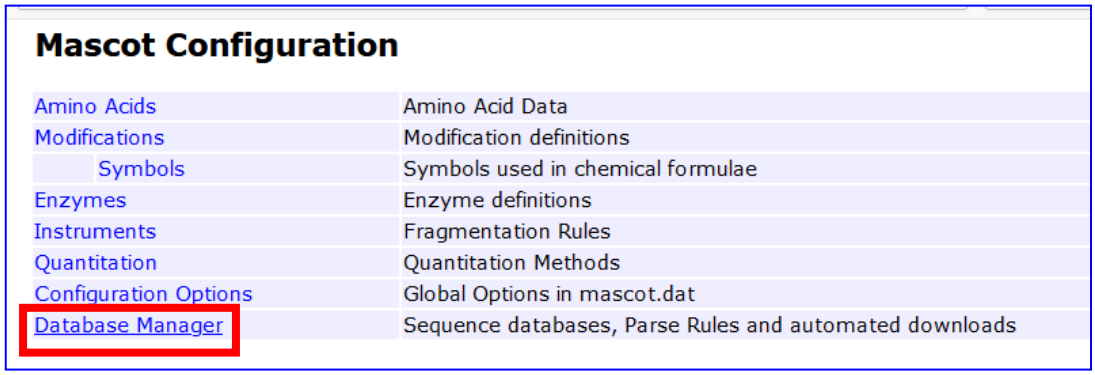
＊選択項目「Compressed/Uncompressed」 は、ダウンロードファイルを圧縮した状態で取得するか、そのまま取得するかの選択肢です。圧縮は「gz」という形式で、圧縮しない状態より大幅にファイルサイズが減少しますが、Windows の標準機能で解凍する事ができません。解凍をするためのソフトウェア（MASCOT のインストール DVD に入っている 7zip など）が利用できる場合は”Compressed”の選択肢をお勧めいたします。Compressed にてファイルを取得した際、**ダウンロード後にファイルを必ず解凍しておいてください。**

＊ **登録件数**が表示されますので、確認しておいてください。データベース登録後に登録件数がダウンロード時の数字と合致するか、チェックする事をお勧めいたします。

C. MASCOT でのデータベース設定

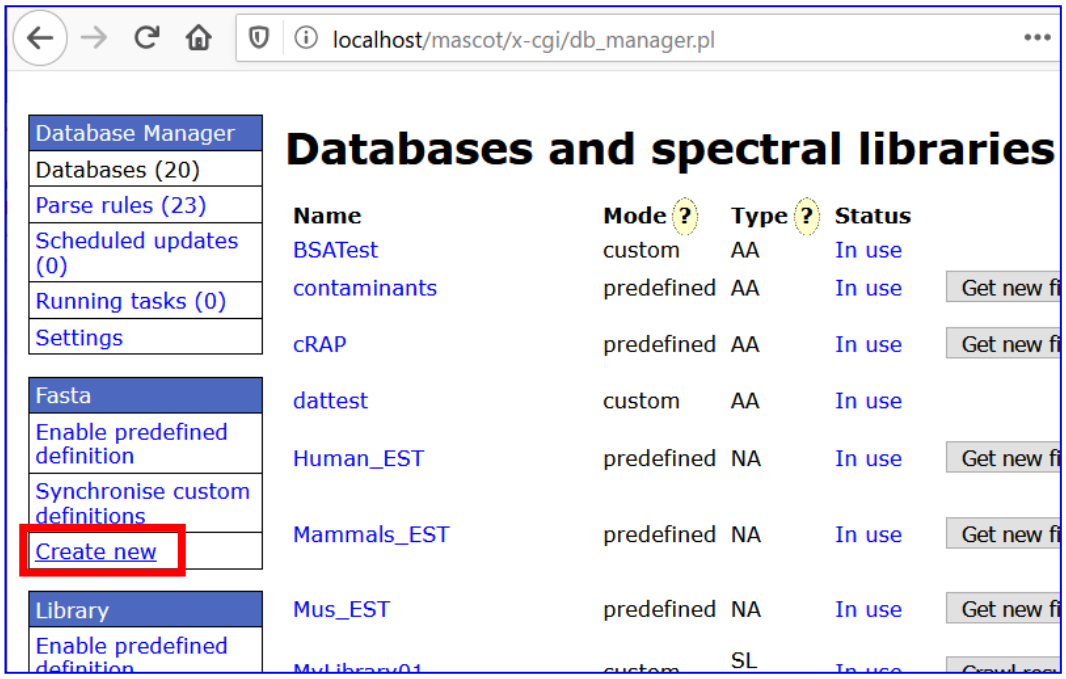
WEB Browser で MASCOT のページへアクセスします。
Home -> **Configuration Editor** -> **Database Manager** とリンクをたどります。



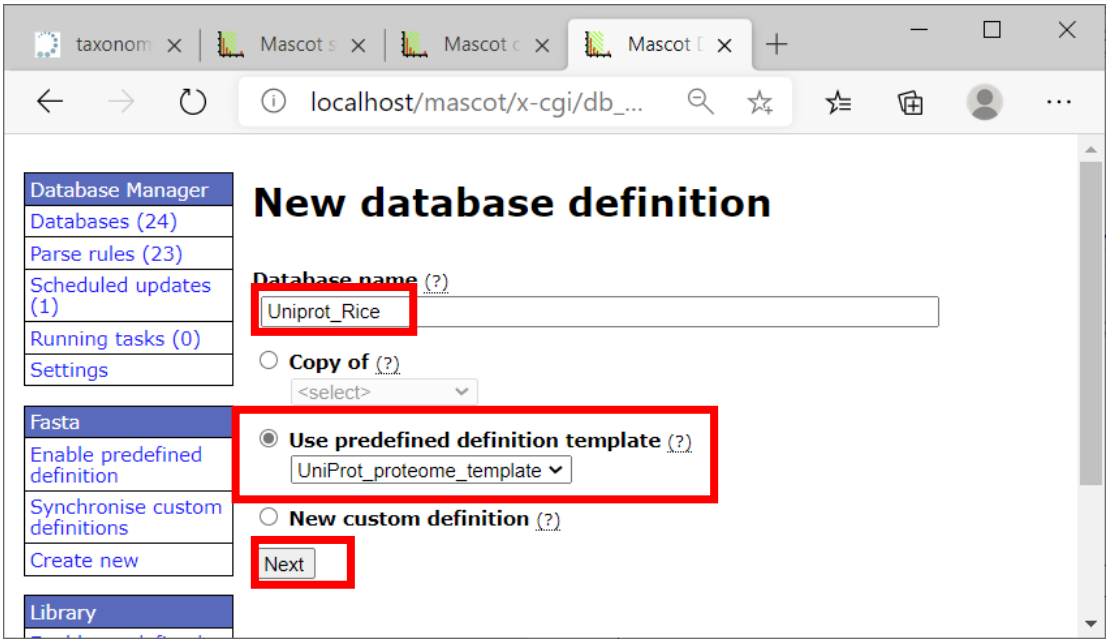


[次頁に続きます]

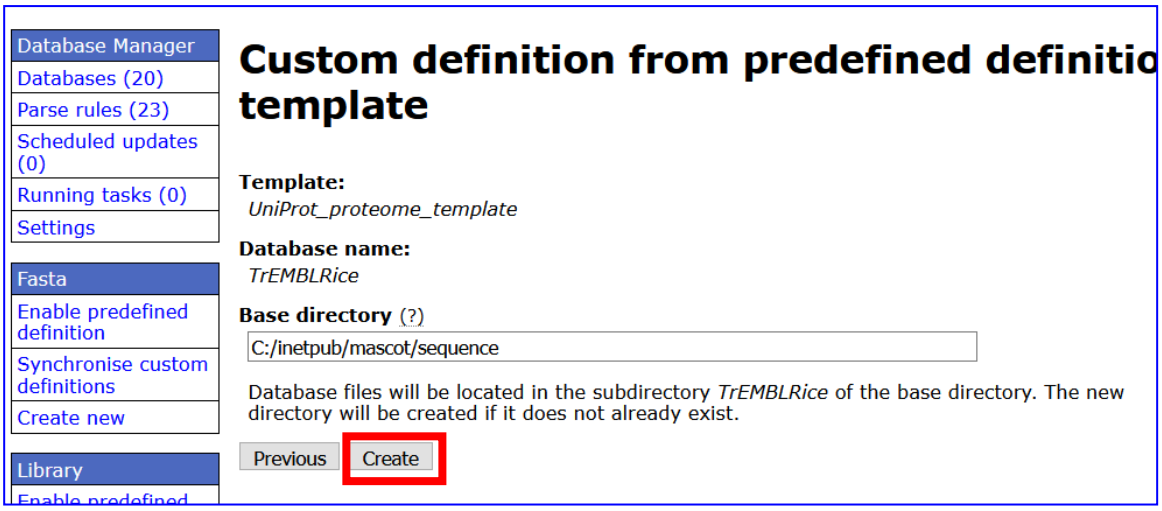
database manager 画面の左フレームにある、「Fasta」の「Create new」をクリックします。



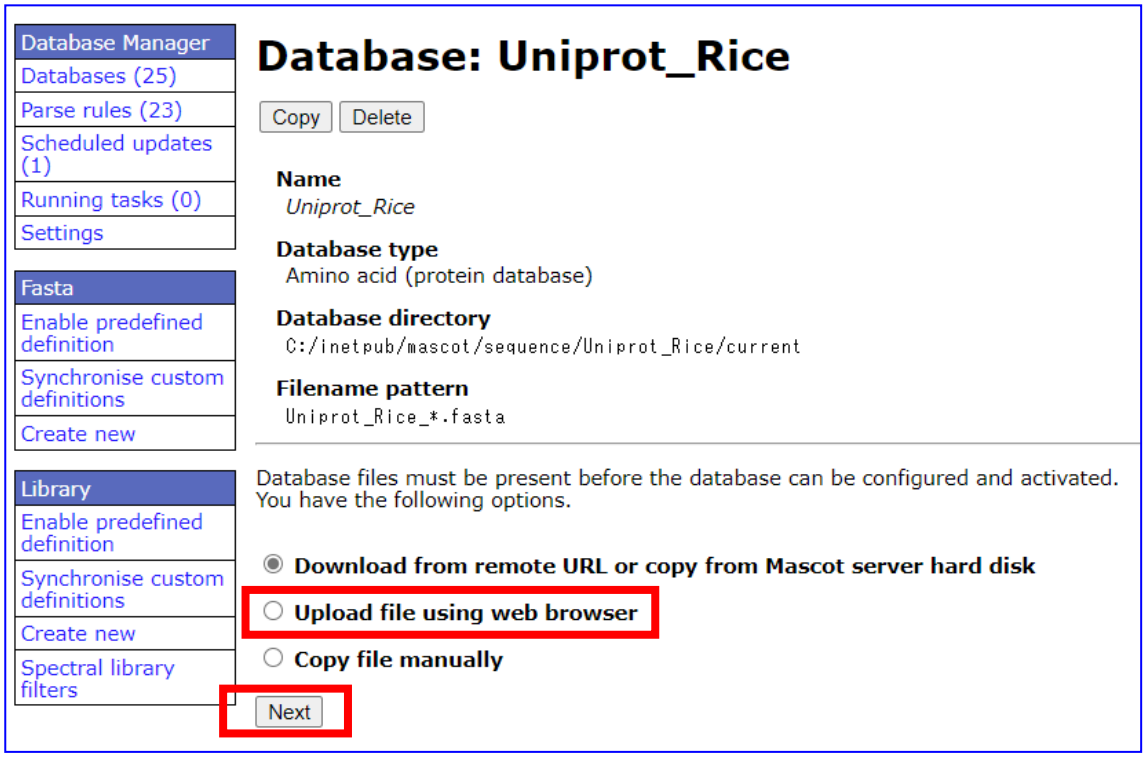
新規設定画面が現れます。「Database name」でデータベースの名称を記入します。その下のラジオボタンでは「Use predefined definition template」選択し、項目として「Uniprot_proteome_template」を選択します。選択後、「Next」ボタンを押します。



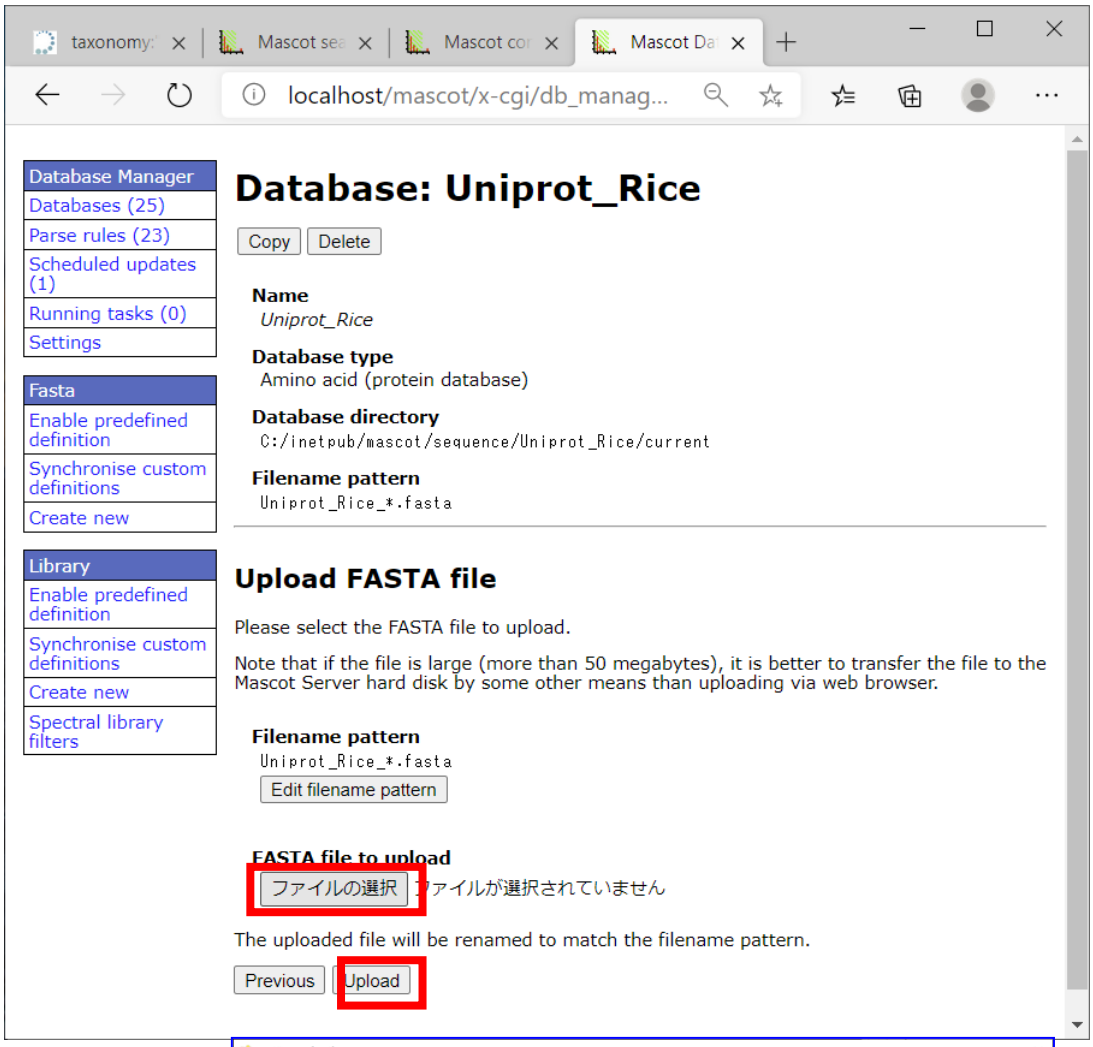
データベースのファイルを置く場所を指定します。記述内容を確認し、「Create」ボタンを押します。MASCOT にてデータベース設定の枠が設置され、所定の場所に配列ファイル並びに関連ファイルを置くためのフォルダが作成されます。

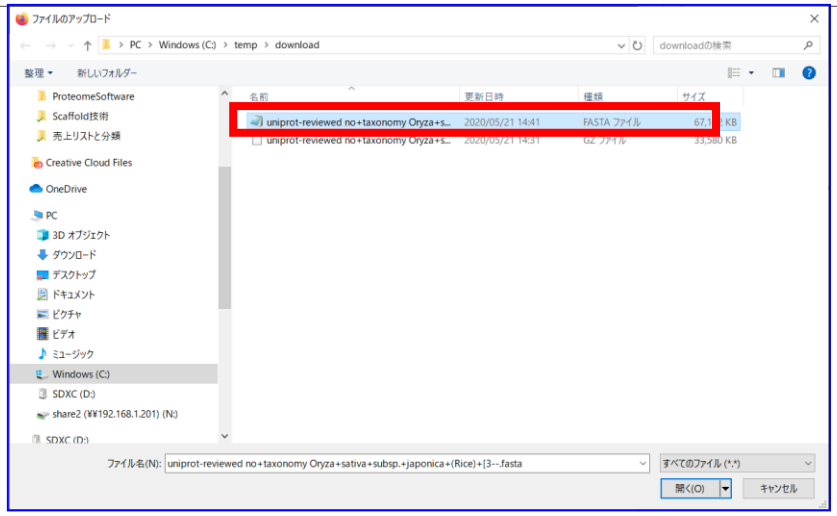


続いてダウンロードしたファイルを MASCOT にセットします。どのような方法でもよいですが、この資料ではアップロードの手法を選択します。「Upload file using web browser」を選択し、「Next」ボタンを押します。

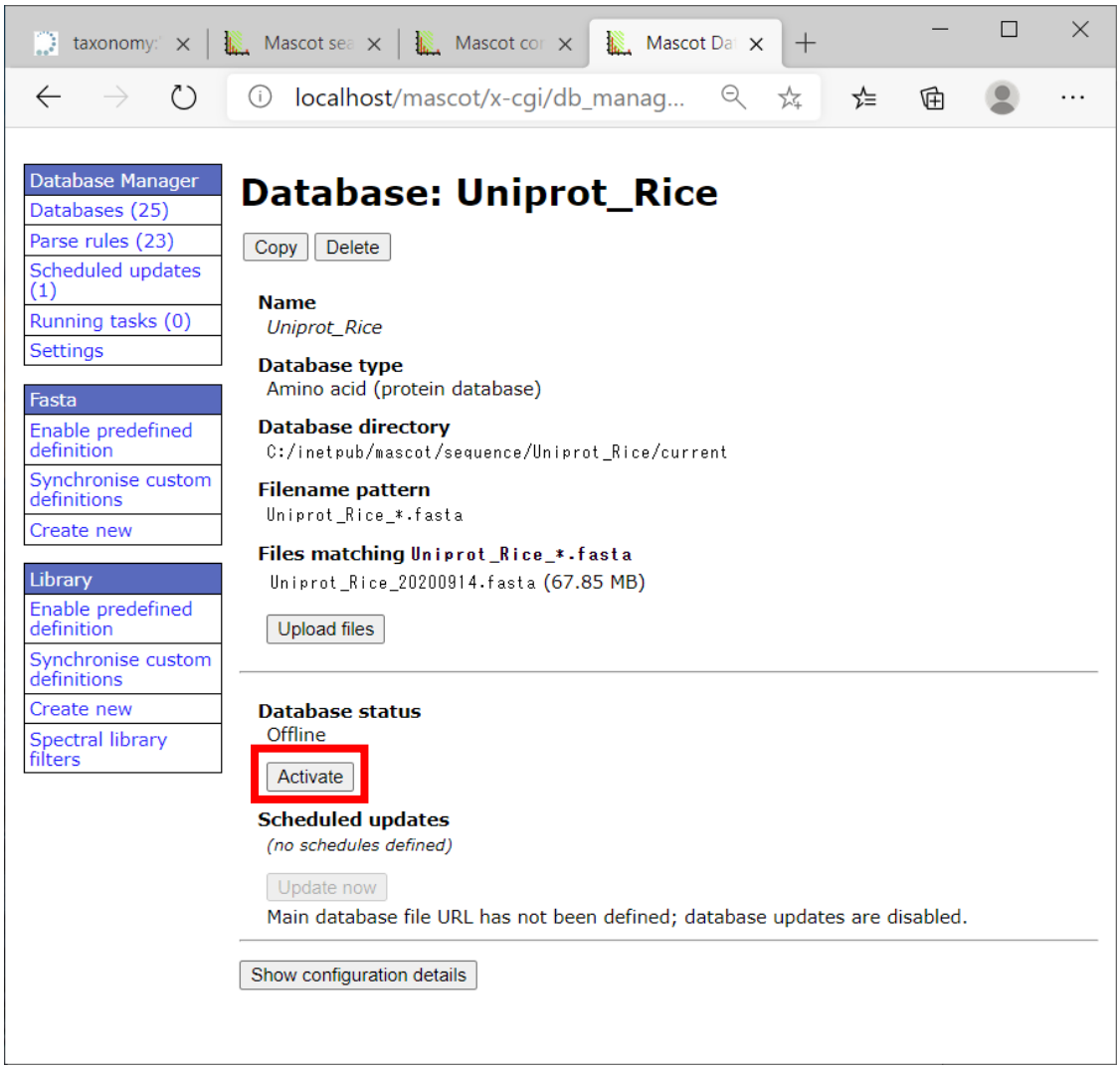


「FASTA file to upload」にある「参照」ボタンを押します。アップロードするファイルとして、Uniprot サイトで取得したファイル（解凍済み）を選択します。ファイル選択し以下の画面に戻った後「Upload」ボタンを押すと、MASCOT がブラウザを介してファイルを所定の場所にコピーします。





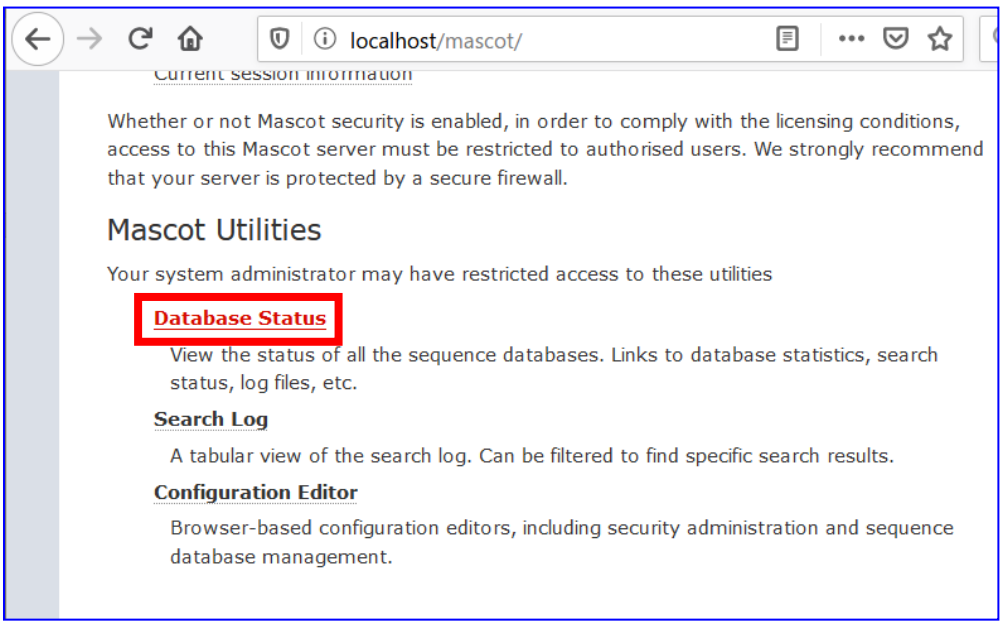
設定を確認する画面が現れます。内容を確認後「**Activate**」ボタンを押すと、データベースの構築が開始されます。

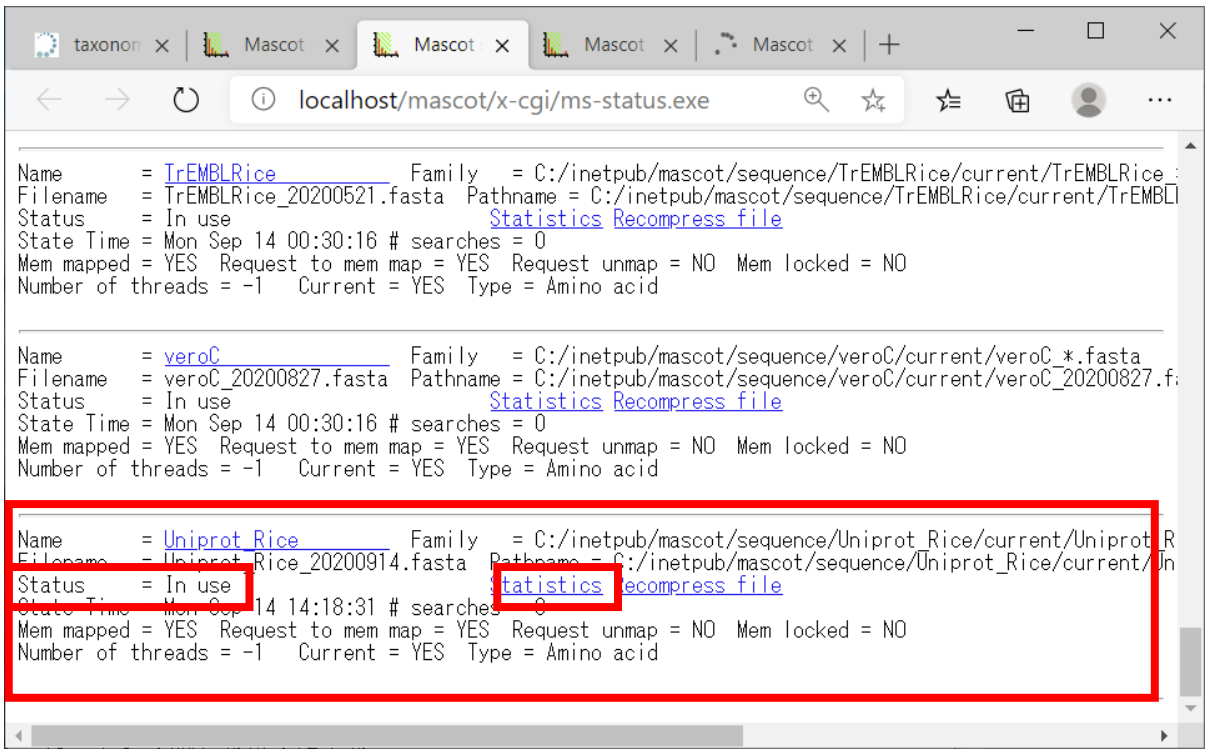


以上で MASCOT におけるデータベースの新規設定は終了です。
データベースの新規設定が終了後、MASCOT では FASTA から MASCOT で使用するデータベース関連ファイルを作成します。その作成が完了後、1 度検索テストを行い正常に検索が完了するかを確認した後、使用可能な状態（status “in use”）となります。

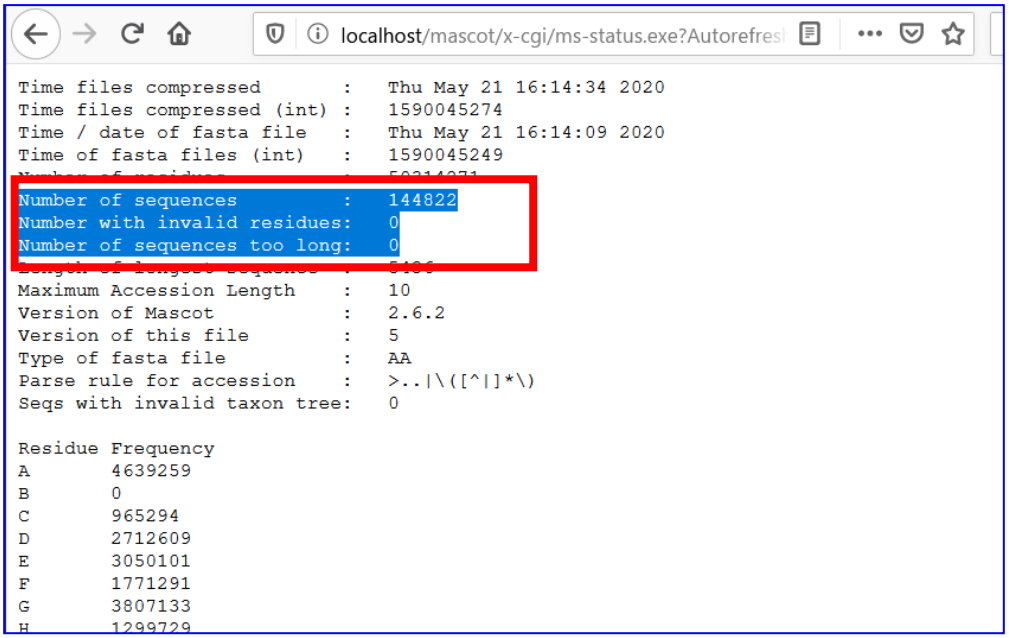
D. データベースの構築確認

データベースの構築状況を確認します。確認は主に Database Status 画面から行います。
Web browser で Database Status (Home -> **Database Status**)画面を開きます。
追加したデータベースが一覧に表示されているか、データベースの”Status”項目が
”**In use**”（使用可能）状態になっているか確認してください。





続いて、”Statistics”のハイパーリンクをクリックしてください。
「Number of sequences」が、ダウンロード時に表示されていたエントリー数と同じか、また「Number with invalid residues」と「Number of sequences too long」が 0 となっているか、確認してください。



データベース構築の確認に関する説明は以上となります。

● 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