NCBInr のフォーマット変更に伴う データベース運用方法の変更

弊社サイトにて公開された以下 URL の情報に基づいた資料です。
http://www.matrixscience.com/help/seq_db_setup_nr.html

［ 本資料内容のまとめ ］

NCBInr のファイルフォーマットが変わりました。gi ナンバーがなくなりました。弊社では、新たなデータベース「NCBIprot」を作成し、「NCBInr」は今後更新せずそのまま置いておく事をお勧めします。

弊社では、nr の新しいフォーマットに対応させた新たなデータベース「NCBIprot」の定義を MASCOT にて提供します。nr の新たな更新ファイルを今後も検索したい場合、本章に記載されている「NCBIprot」の設定方法をご参照の上、データベースを設定してください。

旧「NCBInr」は現在利用可能なバージョンを残したまま置いておくことをお勧めいたします。 2016 年 8 月 21 日 が最後の更新日となります。本章では、以下３つについてご案内いたします。

- 3-1. MASCOT 側で行った NCBInr に対する設定変更内容
- 3-2. nr ファイルがコンピュータに溜まっていないか確認する方法
- 3-3. NCBInr の自動更新を止める方法

特に ver.2.3 以前をご利用の方は必ず自動更新の設定を切るようにしてください。

■ 1. NCBInr のフォーマット変更に伴うデータベース運用方法の変更について

2016 年 8 月 22 日以降にリリースされた NCBInr では、タンパク質エントリのタイトル行の先頭に配置されていた gi 番号（例：489223532）が削除されています。Mascot で使用していたデータベース「NCBInr」は gi 番号をタンパク質エントリーの固有番号として使用していますので、8 月 22 日以降の NCBInr の更新プロセスは完了せず、ダウンロードされた関連ファイルが解凍された直後に停止します。構築に失敗した Fasta ファイルは current フォルダにたまったままの状態となります。

8 月 22 日以降の NCBInr では、gi 番号に代わってアクセッション番号とそのバージョン情報を「.」で結合した文字列がタイトル行の先頭に配置されています（例：WP_003131952.1）。

2016 年 8 月 21 日以前のタンパク質エントリー例

>gi|489223532|ref|WP_003131952.1| 30S ribosomal protein S18 [Lactococcus lactis]
MAQQRRGGFKRRKKVDFIAANKIEVVDYKDTELLKRFISERGKILPRRVTGTSAKNQRKVVNAIKR
ARVMALLPFVAEDQN

2016 年 8 月 22 日以降のタンパク質エントリ例

>WP_003131952.1 30S ribosomal protein S18 [Lactococcus lactis]
MAQQRRGGFKRRKKVDFIAANKIEVVDYKDTELLKRFISERGKILPRRVTGTSAKNQRKVVNAIKR
ARVMALLPFVAEDQN

Mascot がリリースされた 1999 年以来、NCBInr は gi 番号をタンパク質エントリの固有番号として使用していました。NCBInr の固有番号認識に関する設定を変更した場合、過去の検索結果の概要ページは表示されますが Protein View ページなど詳細情報を表示できません。そこで、**2016 年 8 月 21 日以前の NCBInr はそのまま置いておき、2016 年 8 月 22 日以降の NCBInr を新たに「NCBIprot」という名称のデータベースをセットアップ**することをおすすめします。本資料の以降の内容で、新しいデータベース「NCBIprot」の設定方法と、旧データベース「NCBInr」に関する確認事項・設定変更するべき内容についてご案内いたします。

■ 2. 新データベース「NCBIprot」の構築方法

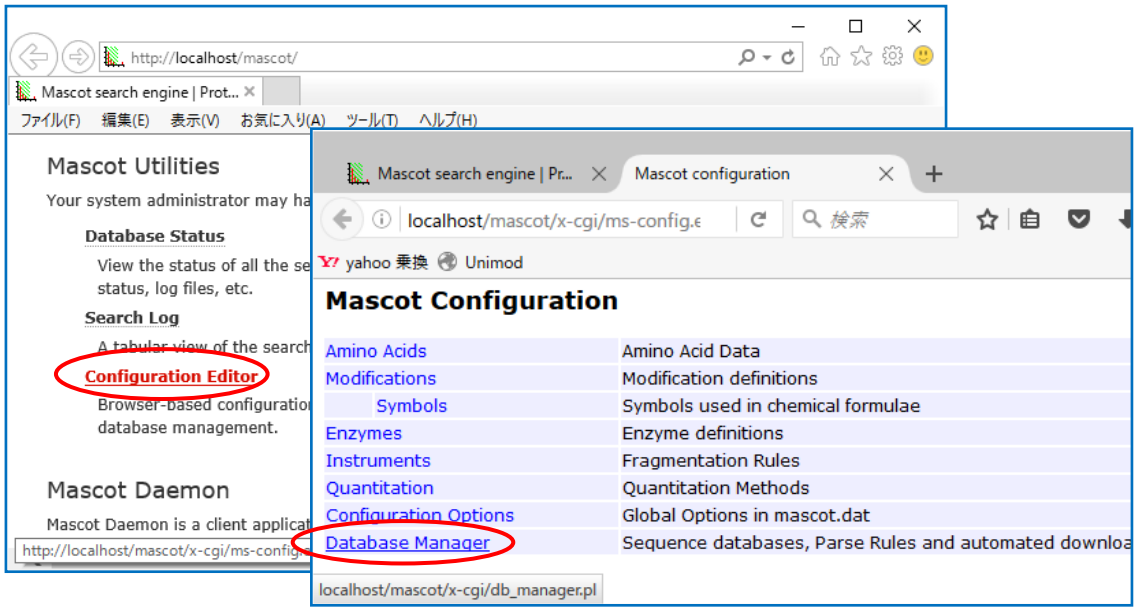
Matrix Science 社では、2016 年8月 22 日以降にリリースされた nr ファイルを使用可能なデータベースの設定として名称「NCBIprot」を準備いたしました。**ver.2.4 以降のユーザーの方にはインターネットを通じて新たな定義「NCBIprot」を自動的に取得することができます。**一方、ver.2.3 以前のバージョンをご利用の方はご自身でデータベースを作成する必要があります。ver.2.3 以前のバージョンにおける NCBIprot の作成手順については、以下弊社手順（英語）をご参照ください。

http://www.matrixscience.com/help/seq_db_setup_nr.html

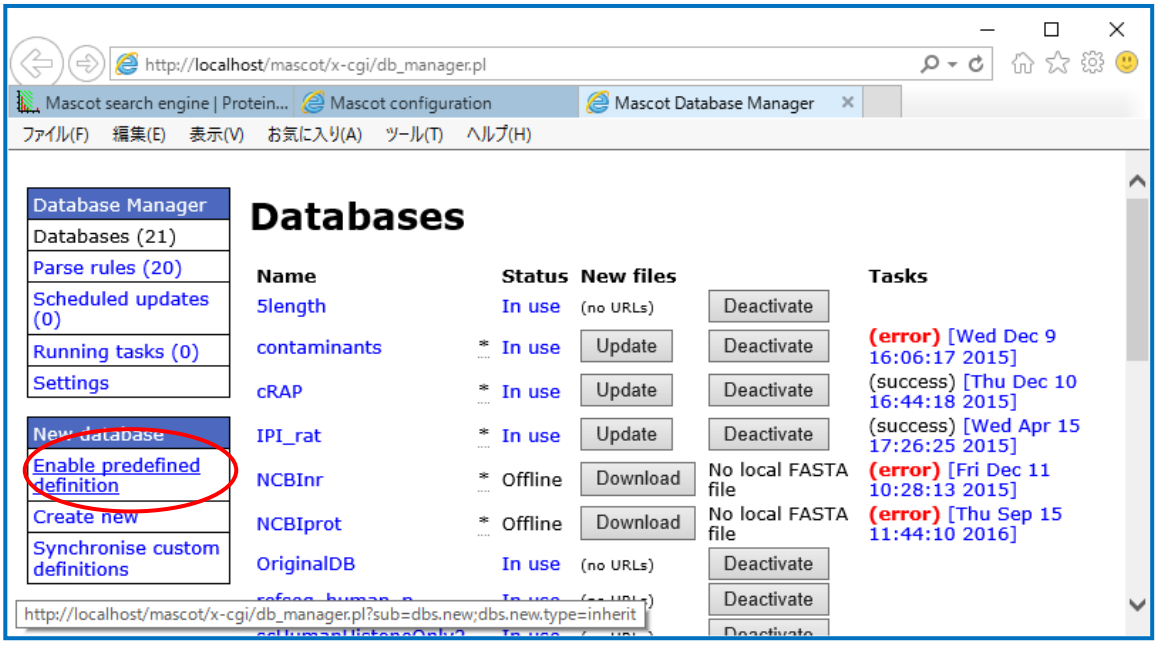
本資料では ver.2.4 以降のユーザーの方向けに、データベース「NCBIprot」を使用可能にする方法について以下ご案内いたします。なお「NCBIprot」の構築には、大きなファイルのダウンロード（nr.gz など、2016 年９月１３日現在、ファイルサイズが 22GB）とファイルの解凍、並びにデータベースの構築と検索テストを行う必要があります。**データベースの構築並びに検索テストには 15～24 時間ほどかかります。**

■ NCBIprot を使用可能にする方法：操作手順

- NCBIprot が定義に含まれているかを確認します。Database Manager 画面を開きます （Home -> Configuration Editor -> Database Manager)。

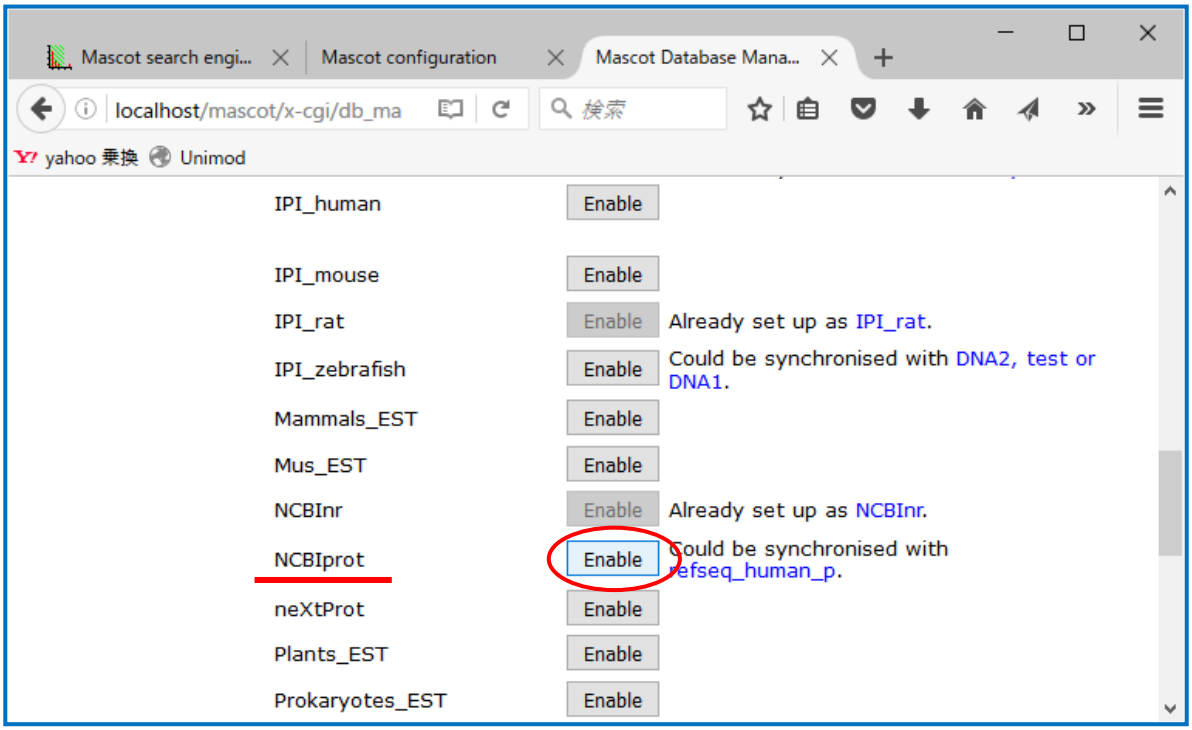


・左のフレームにある「Enable predefined definition」をクリックします。

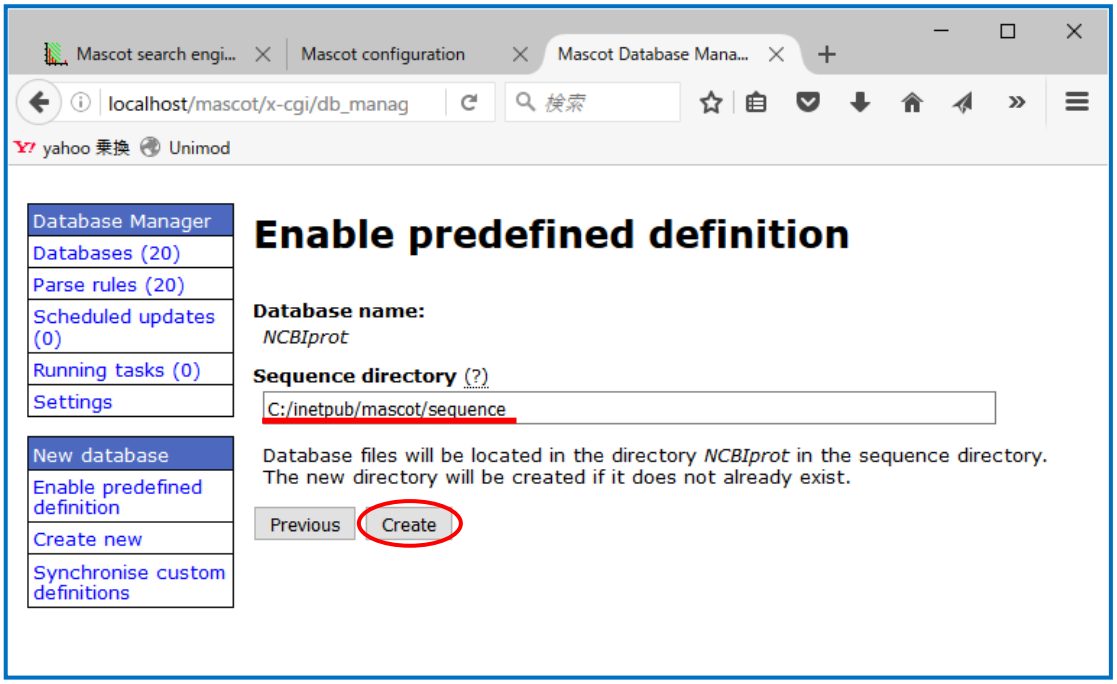


・「NCBIprot」の隣りにある「Enable」ボタンを押します。

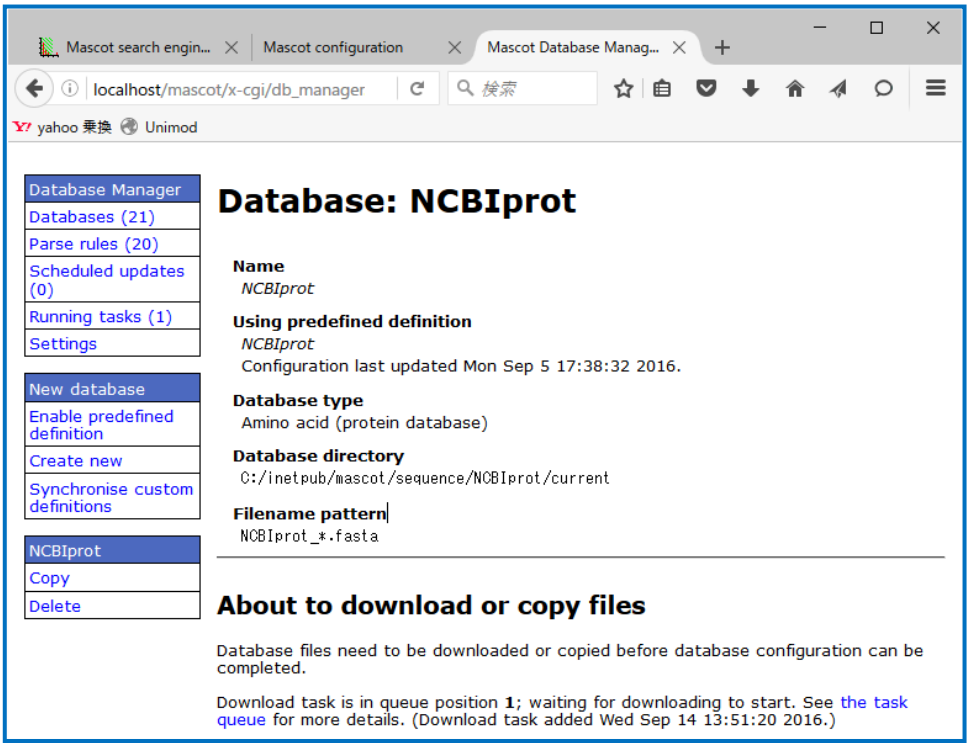
＊この段階で「NCBIprot」の定義が一覧にない方はお手数ですが弊社までご連絡ください。



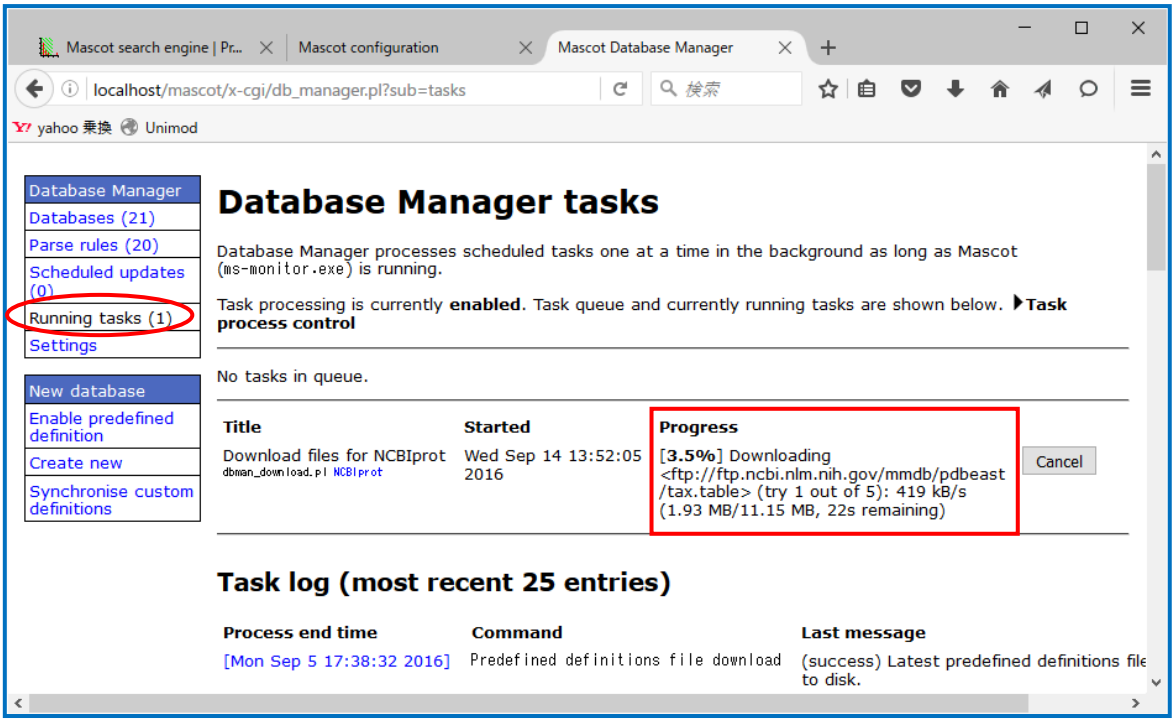
・データベースファイルの置き場所(下記画面では C:/Inetpub/mascot/sequence)を設定した後、「Create」ボタンを押します。



・データベースファイルのダウンロードが始まります。後の処理は基本的にすべて自動で行われます。ファイルサイズが非常に大きい（2016 年 9 月 13 日現在 約 22GB ）ため、ダウンロードに非常に時間がかかります。ご注意ください。



・ダウンロードの進捗を確認するためには、左フレームのリンク「Running tasks」画面をご覧ください。ファイルのダウンロードから解凍までの進捗状況を見ることができます。



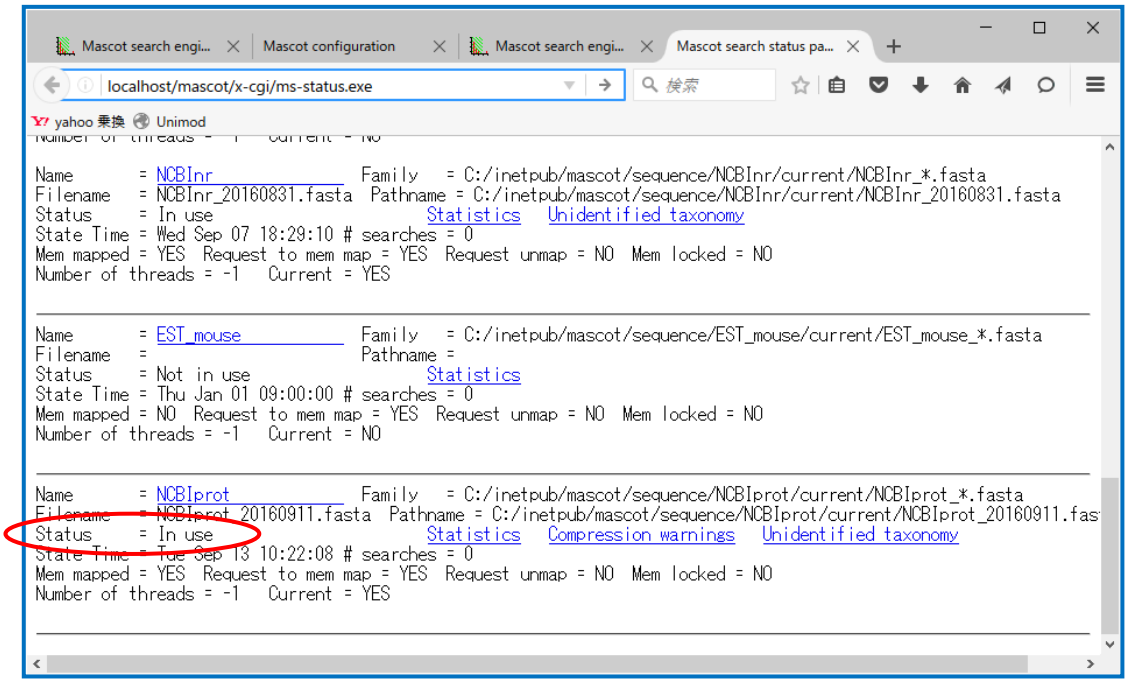
Runnging Tasks のダウンロード進捗を表す画面（上図中段）では、
「Title」に、どのデータベースにおけるファイルダウンロードを行っているか、
「Started」に、作業を開始した時間、
「Progress」にダウンロードの進捗（%表示とダウンロードしたファイルサイズ）
が表示されます。

Database manager の「Running Tasks」画面について、画面の見方に関する詳しい情報は、弊社日本語資料サイトにあります「配列データベース管理マニュアル」
http://www.matrixscience.jp/pdf/jap_database_manager.pdf
の P.43 ～「データベースダウンロード進捗の確認」にもございます。必要に応じてそちらも併せてご覧ください。

・ダウンロード後、データベースの構築が行われます。マシンスペックにもよりますが、構築にも非常に時間がかかる（15～24 時間程度）のでご注意ください。Database Status 画面（Home -> Database Status）にて、「NCBIprot」の status 項目をご覧いただくと構築の進捗状況を確認することができます。 使用可能となるまでの手順は、

データベースの構築→検索テスト→使用可能　　です。

「status」項目が「In use」（使用可能）となれば構築が終了し使用可能となります（下図）。



構築終了後、小さな query データサイズで、taxonomy に何かしらの指定を行った検索テストを行って頂く事をお勧めいたします。

■ 3. 旧データベース「NCBInr」について

データベースに対して固有番号抜き出しルールの変更やデータベースの定義自体の削除や名称変更を行うと、過去に行った検索結果閲覧に不便が生じます。具体的には、結果概要のページ（最初に表示される画面）は閲覧可能であるものの、タンパク質の詳細情報を表示する protein view などタンパク質の配列情報が必要な画面を閲覧できなくなります。そのため弊社では現在利用可能なバージョンで**「NCBInr」を残したまま置いておく**ことをお勧めいたします。NCBInr については 2016 年 8 月 21 日 が最後の更新日となります。

＊これまでほとんど NCBInr を使われていない場合はこれを機に NCBInr を削除していただく事をお勧めいたします。

今後 NCBInr で新しい nr ファイルをダウンロードしても、更新が失敗して current フォルダに fasta ファイルが次々にたまってしまいます。**ver.2.4 以降では自動的にダウンロード先が変更され、問題が起きない様に調整されています。しかし ver.2.3 以前のバージョンをご利用の方は、自動更新を行わないように設定変更が必要**です。

本章では データベース「NCBInr」に関する以下３つの点についてご案内いたします。

■ 3-1. 「NCBInr」の設定変更内容について

ご利用の MASCOT のバージョンが ver.2.4 以降の場合、データベース「NCBInr」の設定が自動的に変更され、**NCBInr の関連ファイル取得先が NCBI から amazon のクラウドサイト（弊社にて使用している領域）に変更**されています。gi ナンバーを含め各種設定は以前と同じです。アマゾンのクラウドサイトに置かれているファイルは、2016 年 8 月 21 日* のバージョンのまま、今後変わる予定がありません。

＊ 弊社にてファイルをアップロードした日が 2016 年 8 月 31 日でした。そのため構築後の Database Status 画面にて NCBInr の filename には日付が 20160831 と表示されますが、実際には NCBI にて 2016 年 8 月 21 日にアップロードされたファイルです

MASCOT のデータベース自動更新では、取得先のファイルに更新がない限りダウンロードをしません。従って ver.2.4 以降の MASCOT では今後 NCBInr においてフォーマットに関する問題は起きないと思われますが、念のため NCBInr の自動更新の設定をオフにしておくことをお勧めします。詳しくは 後述の「3-3. NCBInr 自動更新を停止する方法」をご覧ください。なお、**ver.2.3 以前の MASCOT では NCBInr 更新に関して、自動的な設定変更が行われません。自動更新の設定を止めないとファイルが溜まってしまう可能性がありますので必ず止めてください**。設定変更方法は同じく後述の「3-3. NCBInr 自動更新を停止する方法」をご覧ください。

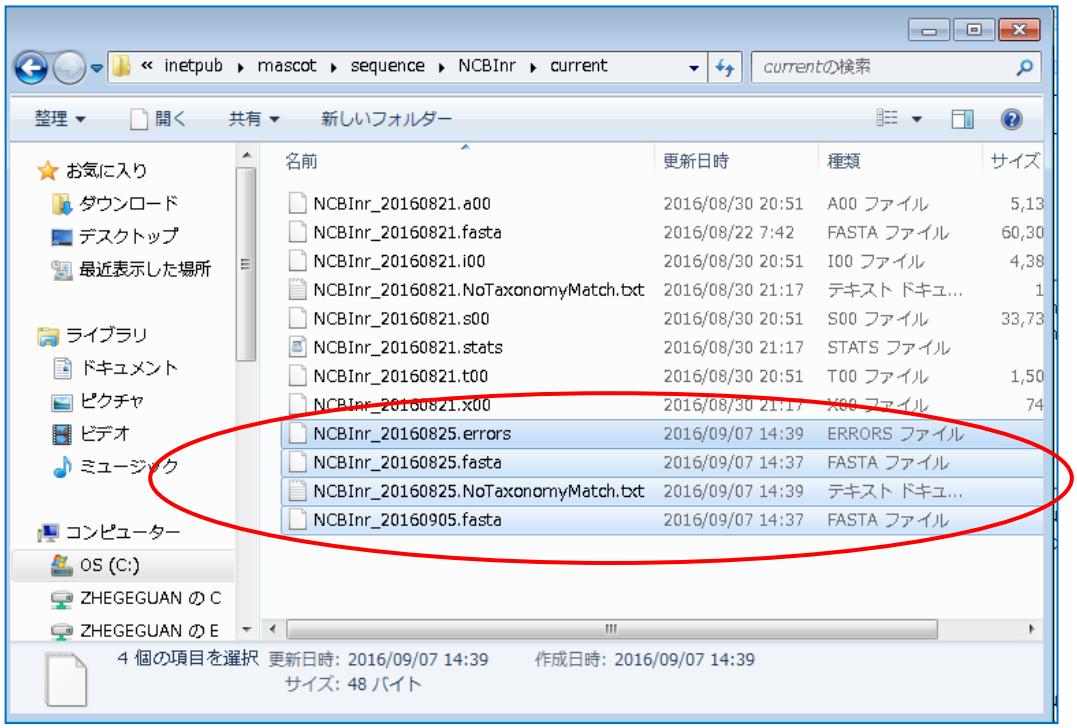
■ 3-2. nr ファイルがコンピュータ内に溜まっていないか確認する方法

フォーマットの変更後に自動更新によってダウンロードされた fasta ファイルは、構築開始直後に失敗して current フォルダ内に溜まります。構築できなかった fasta ファイルは溜まり続けハードディスク容量を圧迫します。**一度 current フォルダの中をご確認いただき、不要なファイルがあれば削除することをお勧めいたします**。以下、その手順ついてご案内いたします。

［操作手順］

C:¥inetpub¥mascot¥sequence¥NCBInr¥current フォルダに不要なファイルが溜まっていないか確認をします。

マイコンピュータを開き、以降 C:¥inetpub¥mascot¥sequence¥NCBInr¥current とフォルダをたどって current フォルダを開きます（下図）。



ファイル名が NCBInr_yyyymmdd.* のファイルが複数存在します。yyyymmdd がファイル更新の日付に該当します。

日付が 2016 年 8 月 22 日以降の fasta ファイル、並びに拡張子が異なる同名の各種ファイルがありましたら、それらをすべて削除してください。削除できない場合、一旦 MASCOT サービスを停止＊1 してから再度同じ操作を試みてファイル削除後サービスを再開＊2 してください。

＊1 サービスを停止する操作
画面左下 Windows マーク→プログラム→MASCOT→admin→stop mascot service
＊2 サービスを開始（再開）する操作
画面左下 Windows マーク→プログラム→MASCOT→admin→start mascot service

■ 3-3. NCBInr 自動更新を停止する方法

NCBInr の自動更新を行う方法は主に２つあり、バージョンにより異なります。ver.2.4 以降では MASCOT のサービス自体に自動更新機能を備えていて、Database Manager でスケジュールを調整しています。一方 ver.2.3 以前ではスクリプトプログラム db_update.pl を、Windows のタスク機能を利用して自動実行させていました。

ver.2.4 以降では NCBInr のファイル取得先に関する設定が自動的に変わり、今後ファイルが更新されることがないため特に大きな問題になりません。しかし ver.2.3 については自動更新を止めないとファイルが溜まってしまいますので必ず自動更新を停止していただくことをお勧めいたします。

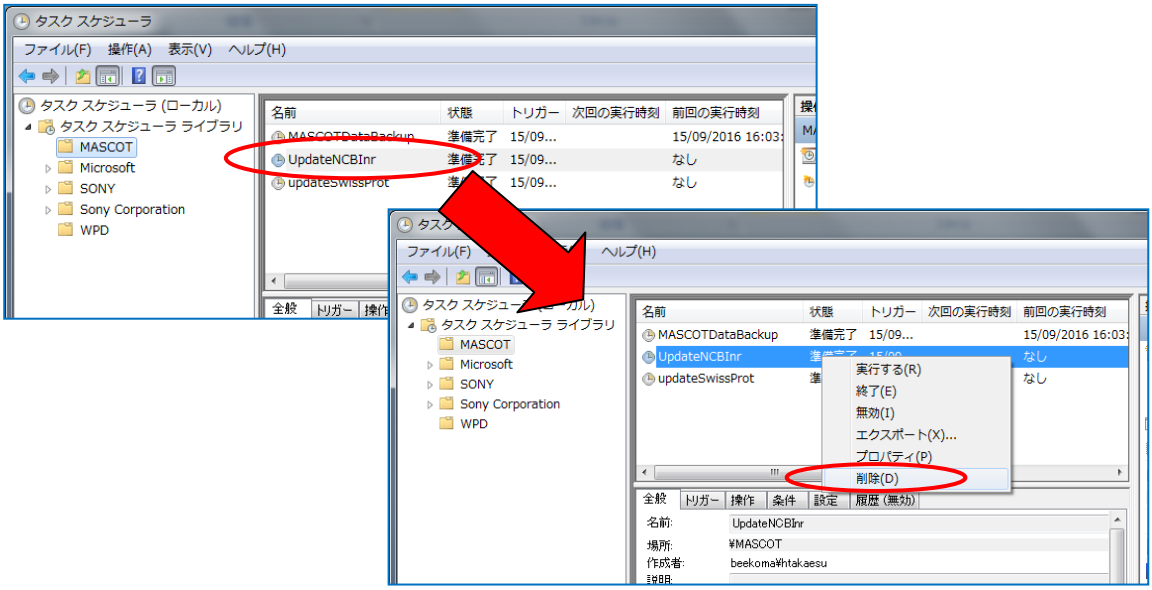
以下、ver.2.3 以前と ver.2.4 以降それぞれで自動更新を止める方法についてご案内いたします。

[自動更新の設定を停止する方法　ver.2.3 以前]

・Windows のタスク設定画面を開きます。
コントロールパネル→システムとセキュリティ→管理ツール→タスクのスケジュール

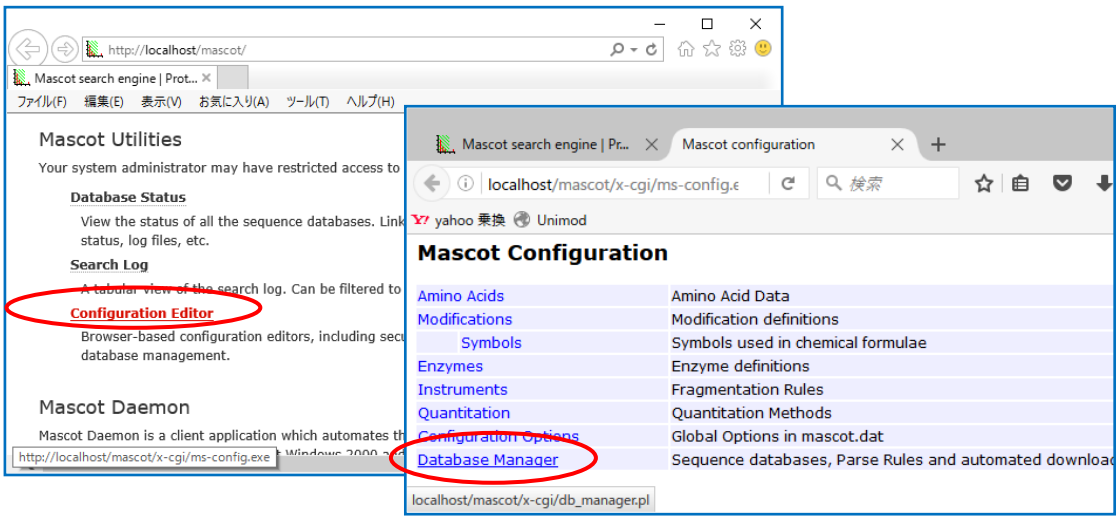
・NCBInr の update に該当する task を削除します。

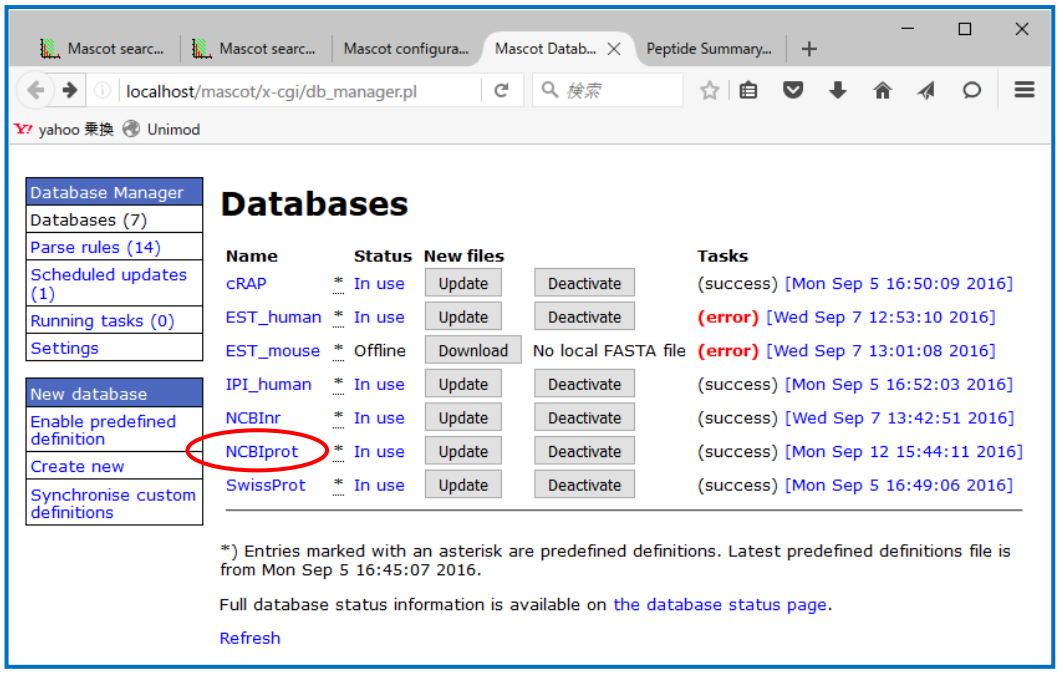
タスクスケジューラー画面にて、
タスクスケジューラーライブラリ → MASCOT → UpdateNCBInr と選び、
右クリック→削除 とします。



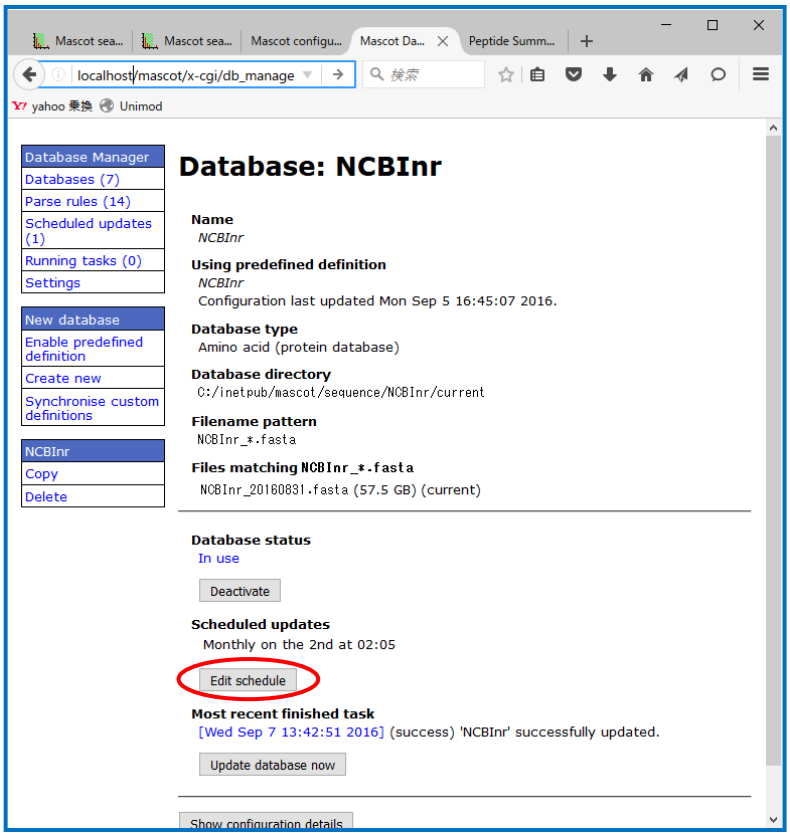
[自動更新の設定をオフにする方法　ver.2.4 以降]

・ Database manager の NCBInr 設定箇所を開きます。
(Home -> Configuration Editor -> Databaase Manager -> NCBInr)

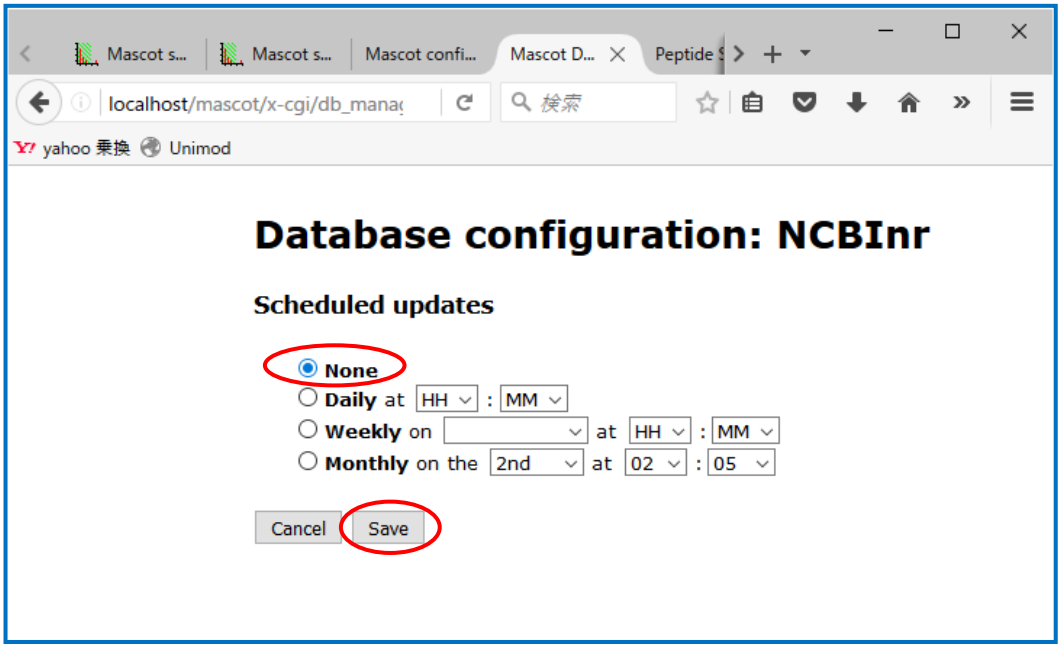




・「Scheduled updates」に既に何かしらの設定がある場合、既に何らかの自動更新設定がされています。削除するにはまず「Edit schedules」ボタンを押します。



・Scheduled updates の選択肢を「None」にして「Save」ボタンを押してください。設定が解除され、自動更新を行わなくなります。



＊ 必要に応じて「NCBIprot」の自動更新設定も行ってください。

ご案内させて頂く操作は以上となります。ご不明な点がございましたらご遠慮なくお問い合わせください。

＜問い合わせ先＞

技術サポート

電子メール　：support-jp@matrixscience.com
電　　話　：03-5807-7897
ファックス　：03-5807-7896