平成１９年度「バイオインフォマティクスＩ」

マルチプルアライメントと分子系統樹基礎

川端　猛

１．演習の準備

まず、次のコマンドを入力して、演習に必要なファイルをコピーして、演習用のディレクトリ **MULTI** に移動してください。

```
cd
cp -r /mandara/lecture/takawaba/MULTI .
cd MULTI
./SETUP
```

※最後のSETUPのコマンドは.cshrcファイルに/mandara/lecture/takawaba/binを加えるためのものです。既に、他のシェルへ変換している方、設定ファイルを自分で変更したい方は、自分で/mandara/lecture/takawaba/binを環境変数PATHに加えてください。

本演習の流れ

演習用のディレクトリに、複数のアミノ酸配列を集めた、配列ファイルがいくつか入っています。これらのそれぞれのファイルに対して、それぞれ以下のステップを実行してもらいます。

(1)解析する配列データの準備
(2)マルチプルアライメントを作成
(3)系統樹を作成
(4)系統樹の表示

2.解析する配列データの準備

以下のような解析対象の複数の配列が入ったFASTA形式のファイルを用意します。今回の演習では、演習用のディレクトリに既に準備してあります。

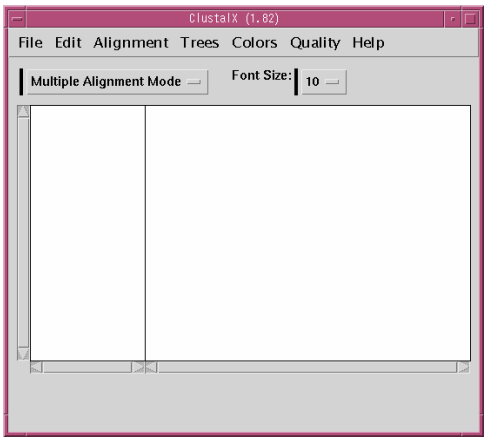
```
>PLAS_ENTPR [P07465] "Plastocyanin"
AAIVKLGGDDGSLAFVPNNITVGAGESIEFINNAGFPHNIVFDEDAVPAGVDADAISAED
YLNSKGQTVVRKLTTPGTYGVYCDPHSGAGMKMTITVQ
>PLAS_ULVAR [P13133] "Plastocyanin"
AQIVKLGGDDGALAFVPSKISVAAGEAIEFVNNAGFPHNIVFDEDAVPAGVDADAISYDD
YLNSKGETVVRKLSTPGVYGVYCEPHAGAGMKMTITVQ
>PLAS_CHLRE [P18068] "Plastocyanin, chloroplast precursor (PC6-2)"
MKATLRAPASRASAVRPVASLKAAAQRVASVAGVSVASLALTLAAHADATVKLGADSGAL
EFVPKTLTIKSGETVNFVNNAGFPHNIVFDEDAIPSGVNADAISRDDYLNAPGETYSVKL
TAAGEYGYYCEPHQGAGMVGKIIVQ
>PLAS_CHLFU [P00300] "Plastocyanin"
DVTVKLGADSGALVFEPSSVTIKAGETVTWVNNAGFPHNIVFDEDEVPSGANAEALSHED
YLNAPGESYSAKFDTAGTYGYFCEPHQGAGMKGTITVQ
```

3. clustalx を用いたマルチプルアライメント

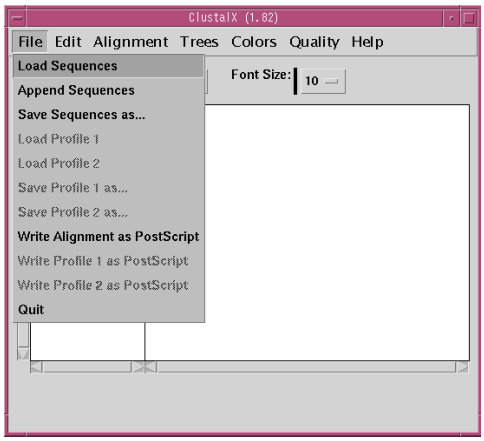
clustalw もしくは clustalx を用いて、行います。まず、clustalx を用いる方法を説明します。
まず、以下のコマンドを入力して、clustalx を実行します。

```
clustalx  &
```

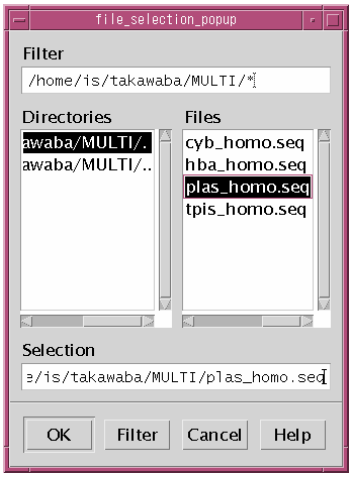
すると以下のウィンドウが立ちあがるはずです。



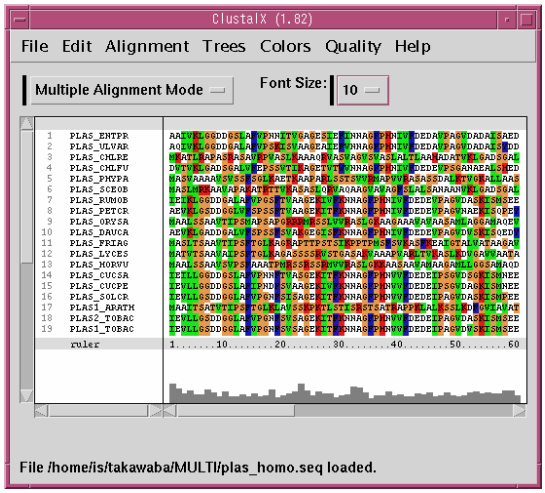
(1) 立ち上げた直後の初期画面はこのようになります



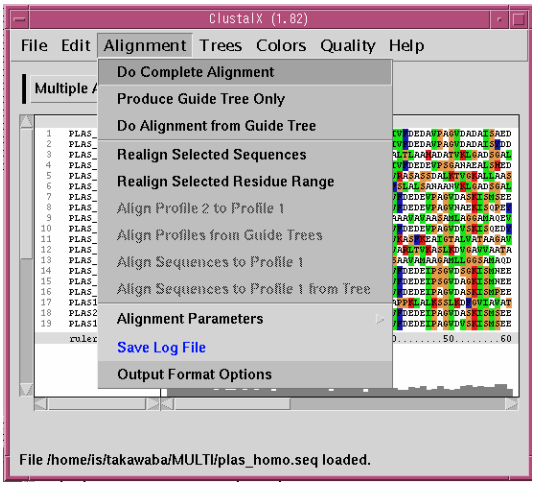
(2) ファイルメニューから、[Load Sequence]を選びます。



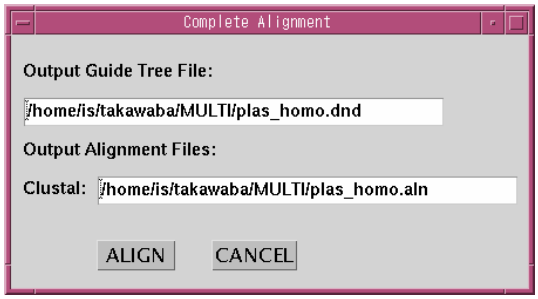
(3) ファイル選択のポップアップが立ち上がるので、解析したい配列ファイルを入力します。



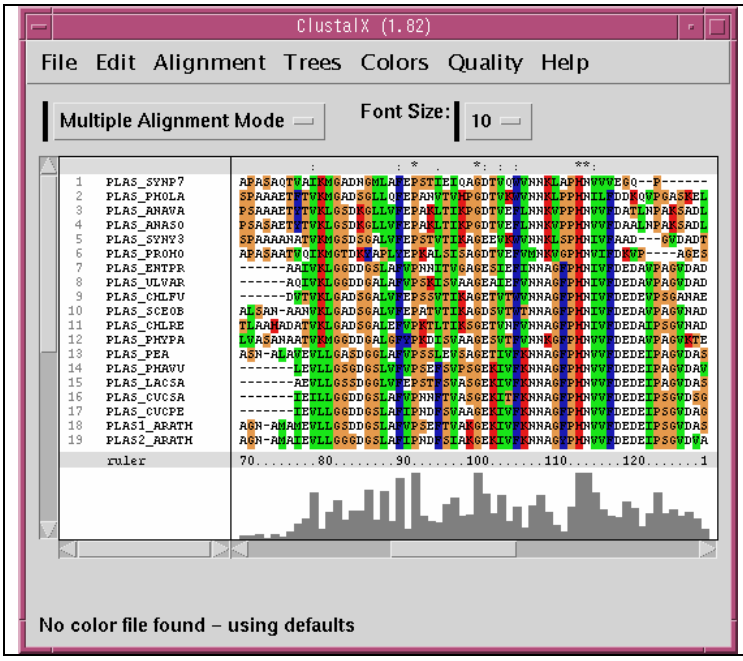
(4) 配列が読み込まれた様子です。まだアライメントはされていません。



(5)[Alignment]メニューから、[Do Complete Alignment]を選びます。



(6)保存するアライメントのファイルを指定します。デフォルトでは、ファイル末尾が**.aln**となるファイルを指定します。



(7)完成したアライメント

マルチプルアライメントの結果は､指定したディレクトリにファイル末尾が aln のファイルとして保存されているはずです。Less コマンドで内容を確認してください。

less ［マルチプルアライメントのファイル名。**plas_homo.aln** など］

以下のようなファイルが得られるはずです。最後のカラムの* : .の記号は、そのサイトの保存の度合いを表しています。

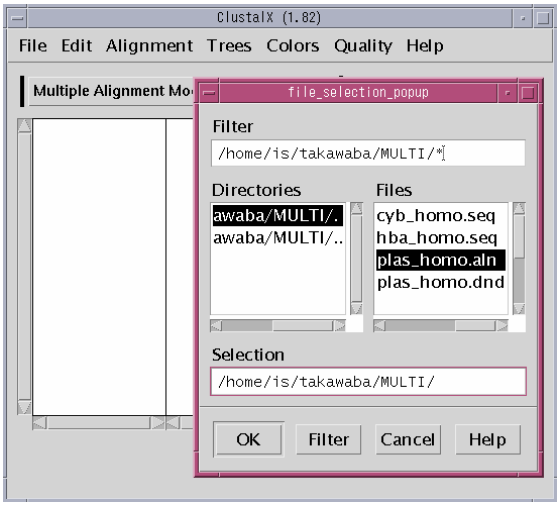
```
CLUSTAL X (1.82) multiple sequence alignment


PLAS_SYNP7      VGSFFLSAAPASAQTVAIKMGADNGMLAFEPSTIEIQAGDTVQWVNNKLAPHNVVVEGQ-
PLAS_CHLFU      --------------DVTVKLGADSGALVFEPSSVTIKAGETVTWVNNAGFPHNIVFDEDE
PLAS_CHLRE      SVAS-LALTLAAHADATVKLGADSGALEFVPKTLTIKSGETVNFVNNAGFPHNIVFDEDA
PLAS_CUCPE      ---------------IEVLLGGDDGSLAFIPNDFSVAAGEKIVFKNNAGFPHNVVFDEDE
PLAS2_ARATH     AAAASIALAGN-AMAIEVLLGGGDGSLAFIPNDFSIAKGEKIVFKNNAGYPHNVVFDEDE
PLAS_SPIOL      ATAAAGLLAGN-AMAVEVLLGGGDGSLAFLPGDFSVASGEEIVFKNNAGFPHNVVFDEDE
PLAS_SOLCR      ---------------IEVLLGSDDGGLAFVPGNFSISAGEKITFKNNAGFPHNVVFDEDE
PLAS2_POPNI     ATAASAMIASN-AMAVDVLLGADDGSLAFVPSEFSVPAGEKIVFKNNAGFPHNVLFDEDA
PLAS_FRIAG      ATAAGAVLASN-ALAVEVLLGGSDGSLAFVPSNIEVAAGETVVFKNNAGFPHNVLFDEDE
PLAS_DAUCA      ---------------AEVKLGADDGALVFSPSSFSVAKGEGISFKNNAGFPHNIVFDEDE
PLAS_HORVU      AMAAGAMLLGGSAMAQDVLLGANGGVLVFEPNDFSVKAGETITFKNNAGYPHNVVFDEDA
AZUP_ALCFA      ILAMLAAPALAENIEVHMLNKGAEGAMVFEPAYIKANPGDTVTFIPVDKG-HNVESIKDM
AZUP_RHILV      ALIASAASLMAADHQVQMLNKGTDGAMVFEPGFLKIAPGDTVTFIPTDKS-HNVETFKGL
                              :         : *  .    *: : :      **:
```

マルチプルアライメントの保存サイトは、一般にその蛋白質の構造や機能を維持するに重要だと考えられます。以下にマルチプルアライメントの保存パターンの意味を考える上で、重要な各アミノ酸ごとの性質の違いを簡単にまとめておきます。

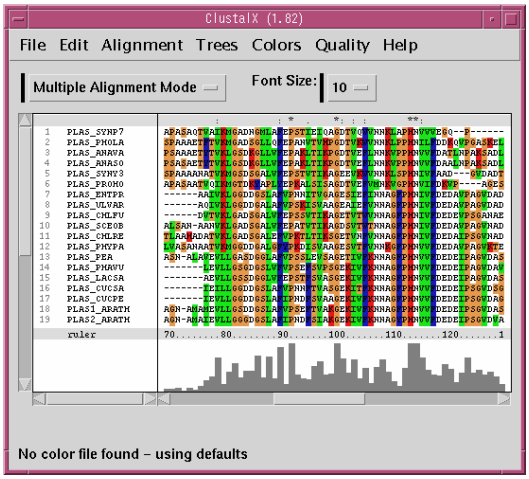
G,P	Glysine は逆巻きヘリックス構造をとることができ、プロリンは構造の逆に構造の自由度が極端に少ないという特徴を持っている。これらの保存は特別な立体構造を維持するのに必要である場合が多い。
H	Histidine は２つの電位状態をとることができるため、酵素の活性部位に現れることが多い。リングを持つ構造を持つため、W,Y,F と同様の役割を果たすこともある。
E,D	負の電荷を持つため、金属イオン等正電荷の分子との結合に重要。活性部位にも現れる。
Q,N	電荷は持っていないが、極性は強く、E,D と似た形状を持つ。活性部位にもよく現れる。
K,R	正の電荷を持っているため、DNA など負電荷の分子との結合に重要である。活性部位にもよく現れる。
C	Cysteine は S-S 結合を形成して、立体構造の安定性に寄与することができる。また、それ以外に Zn や Cu などの金属と配位したり、活性部位として機能する場合もある。
W,Y,F	非極性のリング状の構造を持つ。蛋白質内部の疎水コアの安定性に寄与する一方、糖、核酸、他の蛋白質との結合に重要な役割を果たす場合がある。
L,V,I,A	脂肪族の疎水的アミノ酸は、蛋白質内部の疎水性コアとして働き、活性部位として機能することはまずない。お互いに置換可能である場合が多く、完全に同一のアミノ酸で保存すること少ない。Leu が周期的に保存するロイシンジッパーは例外的なケース。

3. ClustalW を用いた系統樹の作成

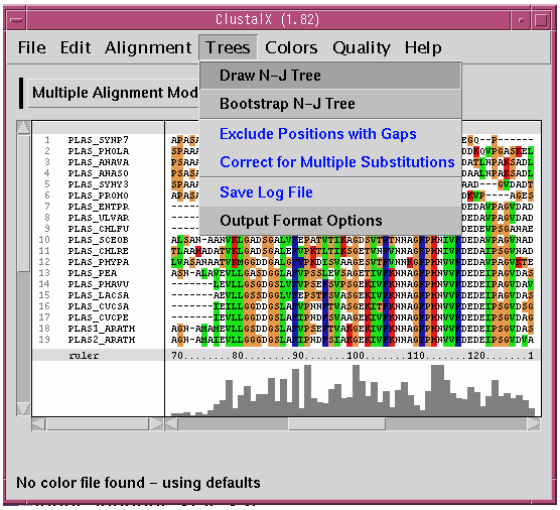
ClustalX にを用いた近隣結合法(N-J 法)による系統樹の作成法を説明します。まず、マルチプルアライメントを読み込みます。もし、マルチプルアライメントを作成した直後に、系統樹を作成するなら、このステップは不要です。



(0)[File]メニューから、[Load Sequence]を選び、マルチプルアライメントファイル（末尾が**.aln** のファイル）を選びます。



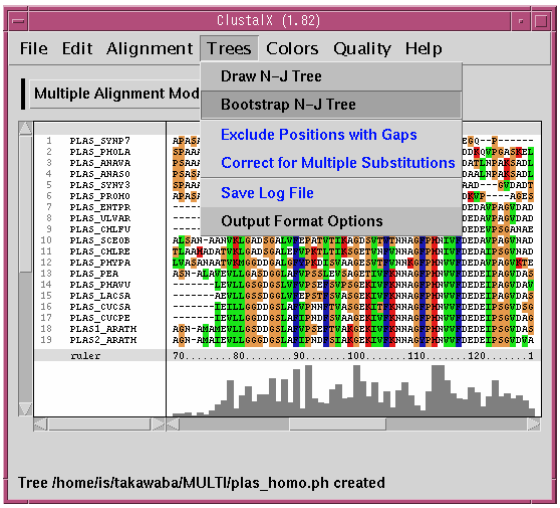
マルチプルアライメントが読み込まれました。



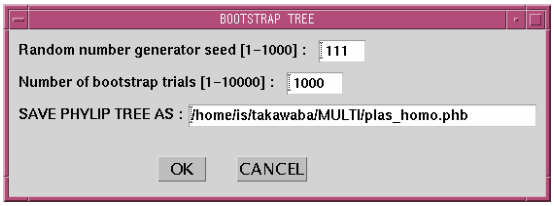
(1)[Trees] メニューから、[Draw N-J Tree]を選びます。

(2)ファイル名を指定します。通常、PHYLIP形式の系統樹のファイルは、ファイル末尾を**.ph** とするのが通例です。

また、同様に N-J 法を用いたブートストラップ値付きの系統樹を計算することも可能です。



(1)[Trees]メニューから、[Bootstrap N-J Tree]を選びます。

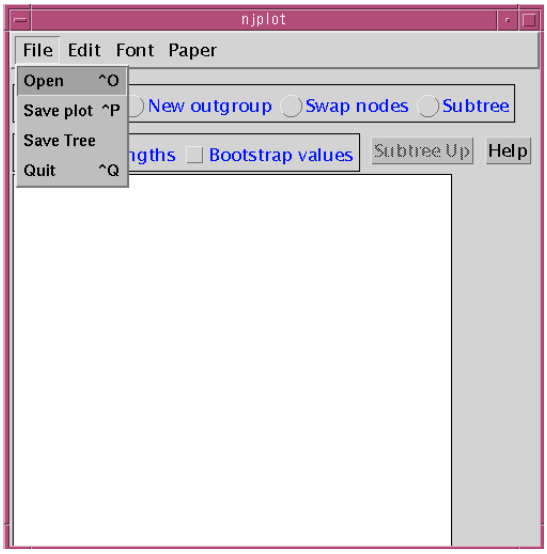


(2)乱数のシード、ブートストラップ標本数とファイル名を指定します。ブートストラップ値付きの PHYLIP 形式の系統樹のファイルは、ファイル末尾を **.phb** とするのが通例です。

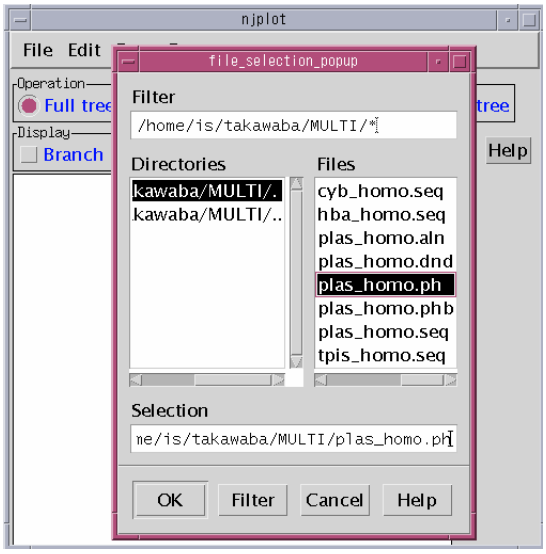
4. njplot を用いた有根系統樹の表示

Njplot は有根系統樹を描画するためのソフトです。

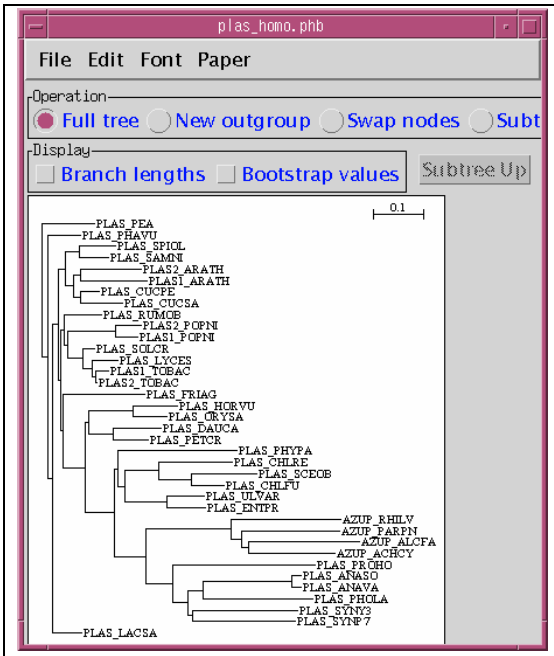
```
njplot &
```



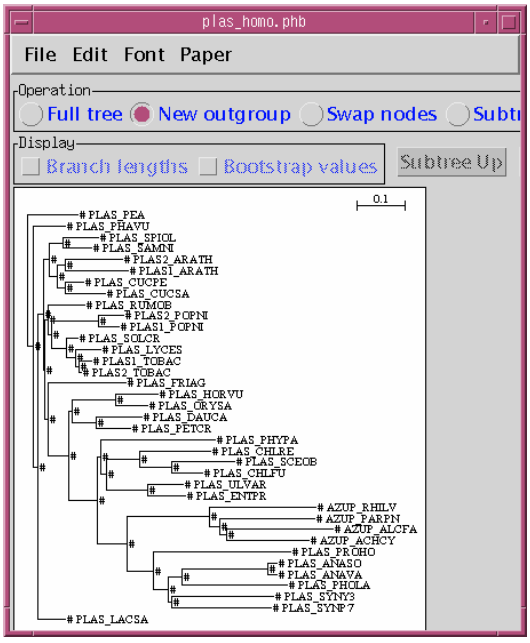
(1)[File]メニューから、[Open]を選びます。



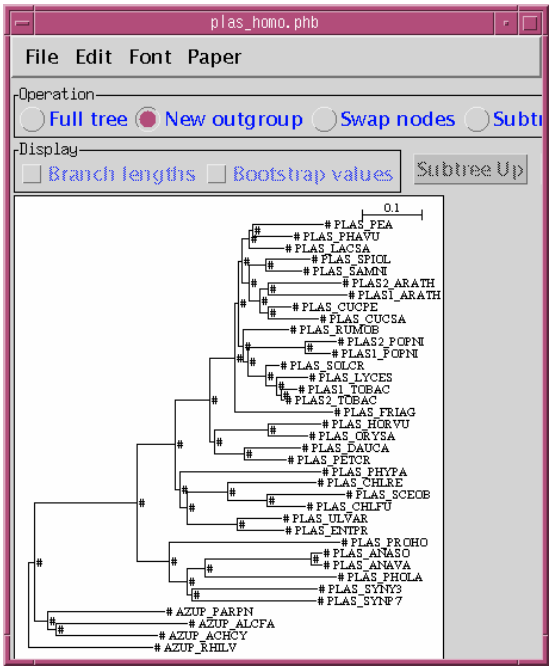
(2)ファイル選択メニューが現れるので、表示したい系統樹のファイル(ファイルの末尾が **.ph** か **.phb** のファイル)を選びます。



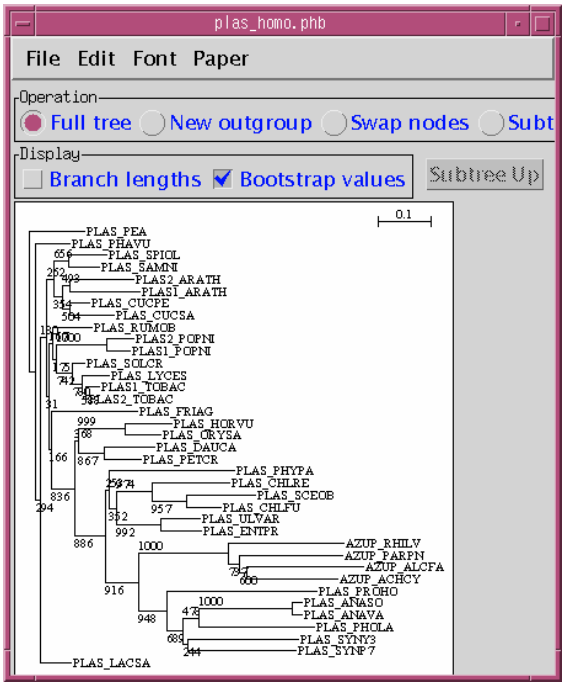
(3)このように有根系統樹が表示されます。ルートの位置は、最も遠い枝長の中点で自動的に決めていますが、必ずしも正しいとは限りません。



(4)[New outgroup]をクリックすると、新しい、外群をマウスで選ぶことができます。



(5)外群を AZUP_RHILV にした場合です。見た目の樹形が大きく変化します。



(6)ブートストラップ値付きの系統樹のファイル(**.phb**)を読み込んだ場合には、[Bootstrap values]をクリックすれば、ブートストラップ値が表示されます。

実習での課題

（1） プラストシアニン・ファミリの配列群をマルチプルアライメントから、保存サイトを確認し、機能上、重要なサイトを推測する。

プラストシアニンは葉緑体の中で、光化学系 I 内の電子伝達を行うタンパク質です。このタンパク質は銅イオンを 1 個結合することが知られています。演習用のディレクトリ、**MULTI** の下に、**plas_homo.seq** というファイルがあり、プラストシアニンファミリの配列が数十本収納しています。これをマルチプルアライメントを行うことで、保存サイトの位置を確かめてください。

保存サイトのパターンを記したものをモチーフといい、それらを集めた ProSite というデータベースがあります。

```
ID COPPER_BLUE; PATTERN.
AC PS00196;
DE Type-1 copper (blue) proteins signature.
PA [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ].
```

ProSite のモチーフの記述は以下のようなルールに従います。

x : 任意のアミノ酸
x(n) : n 個の任意のアミノ酸
x(n,m) : n 個から m 個の任意のアミノ酸
[ACD] : A か C か D のいずれかのアミノ酸
{ACD} : A でも C でも D でもないアミノ酸

さらに保存されるアミノ酸のタイプから、銅結合に関与するサイトを推定してください。金属に配位するアミノ酸は、Cys(C), His(H), Met(M)などが多いとされています。

(2) ミトコンドリアの Cytochrome b の分子系統樹を作成し、それらから、生物種の系統を考える。

ある代表的な遺伝子の分子系統樹を書くことで、その遺伝子の属する生物種の系統樹を推定することができます。これは、伝統的な形態による系統推定に比べて、分子系統学が威力を発揮する問題です。しかしながら、ある遺伝子の分子系統樹が、その遺伝子を持つ生物種の系統樹となるには、**オーソロジー**という関係を持つことが必要です。それはその遺伝子の進化の歴史が、生物種の進化の歴史と同一であることです。それは一見当然のことのように思えますが、実際の進化では、遺伝子重複、遺伝子削除、染色体重複などのイベントが起こるため、一つの生物種に複数の相同遺伝子がある場合があります。よって、単純に相同な遺伝子を集めるだけではオーソロガスにはなっていない可能性があります。また、当然のことながら、対象とする生物種全てが持っている遺伝子でないと系統樹は作成できません。さらに、全般に中立的な進化が行われており、生物種によって極端な進化速度の差がないことも重要な条件となります。

今回の演習では、ミトコンドリアのゲノムにコードされている遺伝子のアミノ酸配列を用います。ミトコンドリアは、酸素呼吸に必要な機能を担うため、全ての真核生物が必ず持っている細胞内器官であること、単純な母性遺伝を行うため、相同組み換えなどのイベ

ントが起こりにくいため、オーソロジーであることがほぼ保証されています。ミトコンドリアのゲノムは小さく、わずか１２個のタンパク質がコードされているだけです。本演習ではそのなかから、Cytochrome b というタンパク質の配列を用います。これはミトコンドリア内膜の電子伝達系の成分で、ユビキノール-シトクローム c 還元酵素というタンパク質複合体の一部を構成します。８本の膜貫通ヘリックスを持ち、２個の HEM を結合します。

演習用のディレクトリは下のファイルが入っています。系統樹は、大きくなるほど計算時間がかかり、意味を読み取るのが難しくなるので、小さなものから順番に試してください。

cyb_mammal.seq	哺乳類の配列群　＋　外群としてニワトリの配列
cyb_reptile.seq	爬虫類の配列群　＋　外群としてサケの配列
cyb_verte.seq	脊椎動物（哺乳類、鳥類、爬虫類、両生類、魚類）の配列群　＋　外群としてハエの配列
cyb_eukary.seq	真核生物（哺乳類、鳥類、爬虫類、両生類、魚類、節足動物、植物、酵母など）の配列群

参考資料　SWISSPROTの５文字表記の生物種名の例

HUMAN Homo sapiens (Human)　ヒト
GORGO Gorilla gorilla gorilla (Lowland gorilla)　ゴリラ
PONPY Pongo pygmaeus (Orangutan)　オラウータン
HYLLA Hylobates lar (Common gibbon)　テナガザル
SAGFU Saguinus fuscicollis (Brown-headed tamarin) タマリン(小型のサル)
MACMU Macaca mulatta (Rhesus macaque) マカクサル
HORSE Equus caballus (Horse)　ウマ
PIG　Sus scrofa (Pig)　ブタ
BOVIN Bos taurus (Bovine)　ウシ
CANFA Canis familiaris (Dog)　イヌ
URSAR Ursus arctos (Brown bear) (Grizzly bear)　クマ
FELCA Felis silvestris catus (Cat)　ネコ
PANTI Panthera tigris (Tiger)　トラ
BALMU Balaenoptera musculus (Blue whale)　クジラ
KOGSI Kogia simus (Dwarf sperm whale)　クジラ
CEPEU Cephalorhynchus eutropia (Black dolphin)　イルカ
ORCOR Orcinus orca (Killer whale)　シャチ
RABIT Oryctolagus cuniculus (Rabbit)　ウサギ
MOUSE Mus musculus (Mouse)　ネズミ
MACGI Macropus giganteus (Eastern gray kangaroo)　カンガルー
SARHA Sarcophilus harrisii (Tasmanian devil)　フクログマ
PHACI Phascolarctos cinerues (Koala) コアラ

LACVV Lacerta vivipara (Common lizard)　トカゲ
PODMU Podarcis mularis (Wall lizard) イワカナヘビ（トカゲの一種）
LACBL Lacerta bilineata (Western green lizard)（トカゲの一種）
IGUIG Iguana iguana (Common iguana)　イグアナ
TERCA Terrapene carolina (Eastern box turtle)　ハコガメ
CHEMY Chelonia mydas (Green sea-turtle) アオウミガメ
APAFE Apalone ferox (Florida softshell turtle)　スッポン
ALLMI Alligator mississippiensis (Mississippi Alligator)　ミシシッピーワニ
ALLSI Alligator sinensis (Chinese alligator) ヨウスコウワニ
CRONI Crocodylus niloticus (Nile crocodile)　ナイルワニ
CAICR Caiman crocodilus (Spectacled caiman)　カイマンワニ
BOACO Boa constrictor (Boa)　ボア（ヘビの一種）
PYTSE Python sebae (African rock python)　パイソン（ヘビの一種）
OPHHA Ophiophagus hannah (King cobra) キングコブラ（ヘビの一種）

CHICK Gallus gallus (Chicken)　ニワトリ
CORBR Corvus brachyrhynchos (American crow)　カラス
VIRLA Vireo latimeri (Puerto rican vireo)　モズモドキ
AQUCH Aquila chrysaetos chrysaetos (Golden eagle)　ワシ

APTPA Aptenodytes patagonicus (King penguin) キングペンギン
EUDCH Eudyptes chrysocome (Rockhopper penguin) イワトビペンギン

XENLA Xenopus laevis (African clawed frog) アフリカツメガエル
RANNI Rana nigromaculata (Japanese pond frog) ウシガエル
RANSI Ranodon sibiricus (Siberian salamander) サンショウウオ

BRARE Brachydanio rerio (Zebrafish) (Danio rerio) ゼブラフィッシュ
SALSA Salmo salar (Atlantic salmon) サケ
SCOSC Scomber scombrus (Atlantic mackerel) サバ
CARAU Carassius auratus (Goldfish) キンギョ
CYPCA Cyprinus carpio (Common carp) コイ
ANGRO Anguilla rostrataSalmo (American eel) ウナギ
LEPSP Lepisosteus spatula (Alligator gar) (Atractosteus spatula)ガー
PRIGL Prionace glauca (Blue shark) サメ
PASSE Pastinachus sephen(Cowtail stingray) エイ

DROME Drosophila melanogaster (Fruit fly) ショウジョウバエ
ANOQU Anopheles quadrimaculatus (Mosquito) カ
ARTSF Artemia sanfranciscana (Brine shrimp) ホウネンエビ

ORYSA Oryza sativa (Rice) コメ
SOLTU Solanum tuberosum (Potato) ジャガイモ
HORVU Hordeum vulgare (Barley) オオムギ
MAIZE Zea mays (Maize) モロコシ
PEA Pisum sativum (Garden pea) マメ
PHYPA Physcomitrella patens (Moss) コケ
LYCES Lycopersicon esculentum (Tomato) トマト
CUCSA Cucumis sativus (Cucumber) キュウリ
ARATH Arabidopsis thaliana (Mouse-ear cress) シロイヌナズナ

YEAST Saccharomyces cerevisiae (Baker's yeast) コウボ