データの用意

アライメントをしやすいデータを整えることがまずは肝要です。自前のデータのみで系統解析する際にはさして気を遣う必要はありませんが、GenBank から落としてきたデータ、特に特定の遺伝子のみの配列データではなくミトゲノムデータなどの場合には適宜必要な部分のみを取り出した方が、この後のアライメントで使うソフトウェアに扱いやすいデータになります。GenBank データからの特定領域の取り出しには GenBank 形式配列から特定遺伝子を切り出すをご参照下さい。

アライメントソフトウェアの選定

基本的には使いやすいソフトで構いませんが、fftnsi(速度重視) か linsi(精度重視) を使うことをお薦めします (どちらも MAFFT に含まれる)。アミノ酸・核酸のどちらでも使えます。

アライメント

データを整えて、MAFFT でアライメントを行う。これが基本です。

データが大きい場合、linsi では時間がかかってしまうので、一旦 fftnsi で大まかにアライメントをした後、BioEdit や MEGA などのアライメントエディタで前後の不要な範囲を除去してから linsi でアライメントを行います。

rRNA 領域などのアライメントが困難な領域では linsi よりも einsi の方が良い結果を得られることがありますのでこちらも試してみて下さい。ただ、私はそもそもアライメントが困難なデータはなるべく使わないようにしています。怪しい部分は捨てるのが原則です。

タンパクコード領域データの場合は、タンパクコード領域塩基配列のマルチプルアライメントにあるように、アミノ酸配列に基づいて行ったアライメント結果と核酸をそのままアライメントした結果を照らし合わせて妥当な方を使います。と言ってもほとんどの場合はその両方式でのアライメント結果が矛盾していたら、その部分は使いません。まれに明らかに最節約的なアライメントがわかることがあるのでそういう場合はそれを使います。

しかし、アライメントは実は系統樹に基づいて行われます。ですから、最終的に導き出された系統樹で同じ結果になるのかを ClustalW で確認します。ClustalW にはアライメント時のガイド系統樹を指定する機能があるので、その際に推定あれた系統樹を与えてやるわけです。もしアライメント結果に変化が生じた場合は再度系統解析を行います。