

Web RIA で始める

バイオインフォマティクス入門

作成 2010年11月 ecobioinfo.com (β版)

はじめに

バイオインフォマティクスとは生物学(biology)と情報科学(informatics)を合成した言葉です。一般的には、生物学を情動的観点から研究すること、生命現象の情報処理、生物学における情報処理技術の応用…とされています。

例えば、DNA から読み取った情報から新しい遺伝子を見つけ出すこと、タンパク発現の解析、遺伝情報からの進化の研究、アミノ酸配列からの生体物質立体構造解析、生体分子間相互作用などコンピュータを活用した生物研究が行われています。そして、これらの情報の共有と有効活用も重要なテーマです。

バイオインフォマティクスと呼ばれる分野が生まれた背景には遺伝学の発展があります。遺伝とは遺伝情報が子孫に伝わっていく現象であるので、情報科学的な観念による研究の対象としては都合の良いものでした。遺伝情報の担い手であるDNA という高分子化合物は ATGC の4種類の塩基の組み合わせによって遺伝情報を保持し、それを複製することによって遺伝情報を子孫に伝えていきます。コンピュータに例えるなら、DNA は4種類の情報の配列を保持する記憶装置で、遺伝とはデータをコピーすることとも言えます。そして、バイオインフォマティクスの初心者はまず遺伝子の解析から始めるのが通例となっていて、古典的な遺伝子解析ソフトウェアである BLAST や ClustalW などは入門者が必ず通った道とも言えます。

現在では Web システムの発達、特に操作性や機能性を重視した RIA(Rich Internet Application) 技術の発達で、特別なソフトウェアや膨大なデータをパソコンにインストールすることなしに様々なツールが使えるようになりつつあります。本書はそのような Web システムを活用したバイオインフォマティクス初心者の為の入門になることを目指しています。

2010年11月 S.Onda

1 必要なもの

Web アプリケーションはパソコンの機種やオペレーティングシステムには依存しませんが性能的には CPU 1GHz 以上、メモリ 512MB 以上、画面の解像度は 1024×768 以上を奨励します。尚、本書では以下のサイトの Web システムを利用します。

国立遺伝学研究所 DNA Data Bank of JAPAN (DDBJ)
<http://www.ddbj.nig.ac.jp/index-j.html>

ecobioinfo.com 遺伝子系統解析 Web システム
<http://www.ecobioinfo.com/>

ブラウザは Microsoft Internet Explorer ver. 6.0 以上、Mozilla Firefox ver. 3.5 以上を奨励します。Flash を使ったシステムを使いますので、ブラウザには Flash Player がインストールされている必要があります。

2 遺伝子情報をデータベースからさがす

日本では国立遺伝学研究所の DDBJ (DNA Data Bank of JAPAN) に全国の研究者から寄せられた遺伝子情報が集められています。米国の GenBank、欧州の EMBL と並んで3大遺伝子データベースと呼ばれ遺伝子データを共有しています。

では、DDBJ にアクセスして遺伝子を探してみましよう。

<http://www.ddbj.nig.ac.jp/intro-j.html>
(現時点では図 2-1 のような画面です。)

DDBJ のサイトには「検索」のメニューがあり、getentry, ARSA, BLAST などの機能が使えます。

Getentry ではアクセッション番号 (accession number) による検索ができます。アクセッション番号とは、登録された遺伝子情報毎に一意につけられた ID です。文献などでアクセッション番号が書いてあればこれで情報を得られます。例えば、アクセッション番号 FJ966983 で検索すると、図 2-2 のような遺伝子情報が表示されます(この表示形式は「フラットファイル」と呼ばれています)。DEFINITION はこの情報の要約で、Influenza A virus (A/Texas/04/2009(H1N1)) segment 7 matrix protein と書いてあることから、2009年にテキサス州で見つかったA型インフルエンザウィルスのマトリクスタンパクの遺伝子であることが解ります。下の方にある ORIGIN が遺伝子の塩基配列情報です(この場合はウィルスの RNA 配列と相補的な

▶ DDBJの紹介

▶ 利用の手引き

▶ Q&A集

塩基配列の登録

- ▶ SAKURA
- ▶ 大量登録システム(MSS)
- ▶ データの修正・更新
- ▶ DDBJ Sequence Read Archive
- ▶ DDBJ Trace Archive

検索

- ▶ getentry
- ▶ ARSA
- ▶ TXSearch
- ▶ BLAST

DDBJ の紹介

DDBJ とは

DDBJ; DNA Data Bank of Japan は、欧州の [EMBL-E International Nucleotide Sequence Database](#) を構成的や国籍に拘わらず閲覧転用していただける世界科
じて INSD にデータを登録することができます。

DDBJ は、[文部科学省](#)からの運営予算で国立遺伝学
す。わが国からの登録の99%以上が、DDBJを通じて

DDBJの事業の柱は、研究者の方々が INSD を使って
規則に従った表現で、できるだけ豊かな情報を記入
することです。

以下の項目別に DDBJ の位置づけをご紹介します

- [塩基配列データベース構築の国際協調体制](#)
- [DDBJ の運営体制](#)
- [DDBJ の主な活動](#)

図 2-1 DDBJ のトップページ

DNA 配列として表記されています。

The screenshot shows the DDBJ (DNA Data Bank of Japan) getentry search interface. The search criteria are: ID 指定: Accession Number, FJ966983. The output format is set to Flat File (DDBJ). The search results are displayed as follows:

Number = [FJ966983]

LOCUS FJ966983 972 bp cRNA linear VRL 01-JUN-2009
DEFINITION Influenza A virus (A/Texas/04/2009(H1N1)) segment 7 matrix protein 2 (M2) and matrix protein 1 (M1) genes, partial cds.
ACCESSION FJ966983
VERSION FJ966983.1
DBLINK Project:37813
KEYWORDS .
SOURCE Influenza A virus (A/Texas/04/2009(H1N1))
ORGANISM [Influenza A virus \(A/Texas/04/2009\(H1N1\)\)](#)
Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
Influenzavirus A

(中略)

ORIGIN

```
1 taaccgaggt cgaaacgtac gttctttcta tcatcccgtc aggccccctc aaagccgaga
61 tcgocagag actggaaagt gtctttgcag gaaagaacac agatcttgag gctctcatgg
121 aatggctaaa gacaagacca atcttgcac ctctgactaa ggaatttta ggatttgtgt
181 tcacgctcac cgtgcccagt gagcaggac tgcagcgtag acgctttgtc caaaatgccc
241 taaatgggaa tggggacccg aacaacatgg atagagcagt taaactatac aagaagctca
301 aaagagaaat aacgttccat ggggccaagg aggtgtcact aagctattca actggtgcac
361 ttgccagttg catgggcctc atatacaaca gcatgggaac agtgaccaca gaagctgctt
421 ttggtctagt gtgtgccact tgtgaacaga ttgctgattc acagcatcgg tctcacagac
481 agatggctac taccaccaat ccactaatca gacatgaaaa cagaatggtg ctggctagca
541 ctacggcaaa ggctatggaa cagatggctg gatcgagtga acaggcagcg gaggccatgg
601 aggttgctaa tcagactagg cagatggtag atgcaatgag aactattggg actcatccta
661 gctccagtgc tggctgaaa gatgacctc ttgaaaattt gcaggcctac cagaagcгаа
721 tgggagtgca gatgcagcga ttcaagtgat cctctcgtca ttgcagcaaa tatcattggg
781 atcttgcacc tgatattgtg gattactgat cgtctttttt tcaaattgat ttatcgtcgc
841 tttaaatacg gtttgaaaag agggccttct acggaaggag tgcctgagtc catgagggaa
901 gaatatcaac aggaacagca gagtctgtg gatgttgacg atggtcattt tgtaacata
961 gagctagagt aa
```

図 2-2 getentry の検索ページと検索結果

ARSA はキーワードから検索できる機能です。例えば、インフルエンザウイルス H1N1 型、テキサス、2009 年で検索してみましよう。ARSA の Quick Search で図 2-3 のように検索条件に Influenza & virus & H1N1 & Texas & 2009 と入力して Search ボタンを押します。

図 2-3 ARSA の検索ページ

図 2-4 の上段ような結果が表示されます。図 2-4 で Sequence Libraries の項目を見ると DDBJ に登録されている検索条件に該当する情報が 1796 件あることがわかります。数字をクリックすると図 2-4 下段のように遺伝子情報の一覧が表示され、アクセッション番号をクリックすると遺伝子情報が表示されます。

Query		Influenza & virus & H1N1 & Texas & 2009	
Sequence Libraries			
DDBJ	1,796	DAD	2,234
UniProt/Swiss-Prot	11	UniProt/TrEMBL	2,177
Sequence Related			
PROSITE	0	PROSITEDOC	0

Primary Accession Number	Definition	Sequence Length
CY044233	Influenza A virus (A/San Antonio/PR921/2009(H1N1)) segment 2 sequence	2,274
CY044234	Influenza A virus (A/San Antonio/PR921/2009(H1N1)) segment 3 sequence	2,151
CY052283	Influenza A virus (A/Texas/43292238/2009(H1N1)) segment 7, complete sequence	987
CY052528	Influenza A virus (A/Texas/45103998/2009(H1N1)) segment 7, complete sequence	987
CY052529	Influenza A virus (A/Texas/45103998/2009(H1N1)) segment 6, complete sequence	1,420

図 2-4 ARSA の結果

BLASTは、入力された遺伝子と類似する遺伝子をデータベースから検索します。通常は未知の遺伝子の塩基配列から類似する遺伝子を検索するのですが、ここでは練習として先ほどのgetentryで獲得した塩基配列情報で検索してみます。この場合の検索条件入力は、「プログラム」の項目はblastn(入力された塩基配列でデータベースの塩基配列から検索)、「検索結果」はwww表示とします。「塩基配列名、検索配列データ」の項目のCOPY&PASTEの欄に前述のgetentryで獲得した塩基配列(ORIGINの部分の塩基配列)をコピーして「入力内容の送信」ボタンを押します(配列の番号などの余計な文字はサーバ側で無効化されます)。

BLAST
version 2.2.24

- プログラム：
 - blastn (塩基配列クエリー × 塩基配列データベース)
 - blastx (塩基配列クエリー[アミノ酸配列に翻訳] × アミノ酸配列データベース)
 - tblastx (塩基配列クエリー[アミノ酸配列に翻訳] × 塩基配列データベース[アミノ酸配列に翻訳])
 - blastp (アミノ酸配列クエリー × アミノ酸配列データベース)
 - tblastn(アミノ酸配列クエリー × 塩基配列データベース[アミノ酸配列に翻訳])
- 検索配列名、検索配列データ：
 - ※検索する配列がひとつの場合は、配列名は必要ありません。ただし配列名をつける場合は、先頭に「>」をつけた配列名を指定してください。
 - ※複数の配列を同時に検索することができます。複数検索の例
 - ※コメントを含んだ検索時指定配列は1MByte以下にしてください。
 - ※デフォルトではフィルターがONになっており、相同性を判断するのにあまり意味がない配列は無効化されます。

File Upload:

or COPY & PASTE:

```

taaccgagst cgaaacgtac gttctttcta tcatcccgtc aggcccccctc aaagccgaga
tcgcgcagag actggaagt gtctttgcag gaagaacac agatcttgag sctctcatgg
aatgctaaa gacaagacca atcttgcac ctctgactaa gggaaattta ggatttgggt
tcacgctcac cgtgccagc gagcggagc tgcagcgtag acgctttgtc caaaatgcc
taaattggaa tggggaccgc aacaacatgg atagagcagc taaactatac aagaagctca
aaasagaaat aacgttccat ggggccaaagg aggtgtcact aagctattca actggtgcac
ttgccagttg catgggcctc atatacaaca ggatgggaac agtgaccaca gaagctgctt
ttggctctag gtgtgccact tttgaacaga ttgctgattc acagcatcgg tctcacagac
agatggctac taccaccaat ccactaatca gacatgaaaa cagaatggtg ctggctagca
ctacgscaaa ggctatggaa cagatggctg gatcgagtgac acaggcagcg gagggcatgg
agtttgctaa tcagactagg cagatggtac atgcaatgag aactattggg actcatccta

```

- 検索結果：
 - WWW Graphical View (<= 100 sequences)
 - ※WWW指定時にGraphical Viewをチェックすると、アラインメントがグラフィカルに表示されます。
 - E-Mail HTML format

- 検索対象データベース：
 - プログラムにより、選択できるデータベースが異なります。
 - 塩基配列データベース
 - DDBJ 全データ (DDBJ 定期リリース + 新着データ)
 - EMBL 新着データ
 - GenBank 新着データ

図 2-5 BLAST の検索ページ

入力内容の送信後は「受付番号は、【.....】です」のメッセージと、入力内容の確認画面となります。「View Result」ボタンを押すと、処理が終了している場合は結果を見ることができます。終わっていない場合は数分待った後に「View Result」ボタンを押してください。処理結果は図 2-6 のようになります。この図では、Influenza A virus H1N1 の遺伝子が類似の遺伝子として検索されているので、入力された遺伝子は A 型インフルエンザ H1N1 の遺伝子であることがわかります。

CLUSTALW SETUP ([Graphical View](#)(≤ 100 sequences) | [Text View](#)(any number of sequences))

BLASTN 2.2.24 [Aug-08-2010]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= query
(972 letters)

Database: /b/DNA.DATA/ddbjhum1.seq;
/b/DNA.DATA/ddbjhum2.seq;
/b/DNA.DATA/ddbjhum3.seq;
/b/DNA.DATA/ddbjhum4.seq;
/b/DNA.DATA/ddbjhum5.seq;
/b/DNA.DATA/ddbjhum6.seq;
/b/DDBJNEW.DATA/new_ddbjhum.seq;
/b/DNA.DATA/ddbjpri1.seq;
/b/DNA.DATA/ddbjpri2.seq;
/b/DDBJNEW.DATA/new_ddbjpri.seq;
/b/DNA.DATA/ddbjrod1.seq;
/b/DNA.DATA/ddbjrod2.seq;

(中略)

Sequences producing significant alignments:			Score	E
			(bits)	Value
GQ894903	GQ894903.1	Influenza A virus (A/Oklahoma/01/2009(H1N1))...	1885	0.0
GQ894856	GQ894856.1	Influenza A virus (A/North Carolina/05/2009(...	1885	0.0
GQ457485	GQ457485.1	Influenza A virus (A/Texas/05/2009(H1N1)) se...	1885	0.0
GQ457473	GQ457473.1	Influenza A virus (A/Kansas/02/2009(H1N1)) s...	1885	0.0
GQ377073	GQ377073.1	Influenza A virus (A/Texas/04/2009(H1N1)) se...	1885	0.0
GQ323514	GQ323514.1	Influenza A virus (A/Texas/10/2009(H1N1)) se...	1885	0.0
GQ232074	GQ232074.1	Influenza A virus (A/Texas/11/2009(H1N1)) se...	1885	0.0
GQ221800	GQ221800.1	Influenza A virus (A/Texas/04/2009(H1N1)) se...	1885	0.0
GQ200220	GQ200220.1	Influenza A virus (A/Texas/12/2009(H1N1)) se...	1885	0.0
GQ162192	GQ162192.1	Influenza A virus (A/Mexico/4482/2009(H1N1))...	1885	0.0
GQ160573	GQ160573.1	Influenza A virus (A/Texas/22/2009(H1N1)) se...	1885	0.0
GQ117050	GQ117050.1	Influenza A virus (A/Texas/08/2009(H1N1)) se...	1885	0.0
GQ117031	GQ117031.1	Influenza A virus (A/Texas/09/2009(H1N1)) se...	1885	0.0
FJ981617	FJ981617.1	Influenza A virus (A/Texas/04/2009(H1N1)) se...	1885	0.0
FJ981608	FJ981608.1	Influenza A virus (A/Texas/05/2009(H1N1)) se...	1885	0.0
FJ966983	FJ966983.1	Influenza A virus (A/Texas/04/2009(H1N1)) se...	1885	0.0
FJ966968	FJ966968.1	Influenza A virus (A/Texas/05/2009(H1N1)) se...	1885	0.0
CY044254	CY044254.1	Influenza A virus (A/San Antonio/PR923/2009(...	1885	0.0
CY044246	CY044246.1	Influenza A virus (A/San Antonio/PR922/2009(...	1885	0.0
CY044238	CY044238.1	Influenza A virus (A/San Antonio/PR921/2009(...	1885	0.0

図 2-6 BLAST の検索結果

3 遺伝子相同性検索を微生物の検査に応用する

類似の遺伝子を検索すること(相同性検索)は、細菌やウイルスなどの微生物を調べるのにも利用できます。目には見えない微生物の種類を調べることは難しいので、遺伝子を利用した方法は比較的簡単で実用的な方法と考えられます。例えば、細菌の場合 16SrDNA と呼ばれる遺伝子は同じ種類の細菌では塩基配列の変異が少ないので種類を判別するのに使われることもあります。例として、ecobioinfo.com の遺伝子系統解析 Web システムを利用した方法をあげます。

<http://ecobioinfo.com/> の「ソフトウェア」のページ

「[遺伝子相同性検索 デモ版はここから起動](#)」をクリック

(本稿執筆時点では作成途中の暫定公開です)

システムを起動すると図 3-1 のような画面が表示されます。「新規登録」ボタンを押して任意の半角英数字のユーザ ID とパスワードを入力してください。登録が完了するとメイン画面が使えるようになります。

図 3-1 遺伝子系統解析 Web システムのメイン画面

ここで、表 3-1 に示す「ある細菌」の 16SrDNA 遺伝子の塩基配列を入力して、この細菌がどのような細菌か調べてみましょう。この配列をコピーして「検索配列データ」の入力域に貼り付けてください(この塩基配列はホームページからダウンロードできますので、ダウンロードしたファイルを読み込んでも構いません)。「検索は配列名」には適当な英数字を30文字以下で入力してください。

(尚、デモバージョンなので「検索対象データベースは」バクテリア 16SrDNA しかありません。)

```
aacacatgcaagtcgaacggtgacgaggagcttgctcctccgatcagtggcgaacgggtg
agtaacacgtgagtaacctgccccagactctggaataacagttggaacagctgctaata
ccggatacagagacggagagggcatctctaccgtctggaaagtttttcggtctgggatggac
tcgcgccctatcagcttggtgaggtagtggtcaccaaggcgacgacgggtagccgg
cctgagagggcgaccggccacactgggactgagacacggcccagactcctacgggaggca
gcagtggggaatattgcacaatgggcgaaagcctgatgcagcaacgccgctgagggatg
acggccttcgggttgtaaacctcttcagtagggaagaagcgaaagtacggctacctaca
gaagaagcaccggctaactacgtgccagcagccggttaatacgtagggtgagagcgttg
tccggaattattgggcgtaaagagctttagggcggtttgtcgctctgctgtgaaaatcc
ggggctcaaccccgacttgagtggttacgggcagactagagtgtggtaggggagactg
gaattcctggtgtagcgggtgaaatgcccagatatacaggaggaacaccgatggcgaaggca
ggtctctgggcccactactgacgctgagaagcgaaagcgtggggagcgaacaggattagat
accctggtagtccacgccgtaaacgttgggaactaggtgtgggtctcattccacgagatc
cgtgccgcagctaacgcattaagttccccgcctggggagtacggccgcaaggctaaaact
caaaggaattgacggggggcccgcaaacgcccggagcatgtggattaattcgatgcaacg
cgaagaaccttaccaggcttgacataataccggaaacaccagagatgggtgccccgcaa
ggtcgggtatacaggtggtgcatggttgctgctcagctcgtgctgagatggtgggttaag
tcccgcacagagcgaacctcgttctatggtgcccagcgtaaaggcggggactcatag
gagactgccggggtcaactcggaggaaggtggggatgacgtcaaatcatcatgccctta
tgtcttgggcttcacacatgctacaatggccggtacaaagggctgcgaaatcgcgagatg
gagcgaatcccaaaaaaccggtctcagttoggattggggtotgcaactcgaccccatgaa
gtcggagtcgctagtaatcgcagatcagcaatgctgcgggtgaatacgttcccgggccttg
tacacaccgccgctcaagtcacgaaagtcggtaaacaccggaagccggtggcccgaacct
tgtgggggga
```

表 3-1 サンプルの DNA 塩基配列

遺伝子相同性検索の方法(検索のアルゴリズム)には、前述の BLAST や FastA, Ssearch など幾つかの方法があります。Ssearch は生物学的に厳密な結果を得られる Smith-Waterman アルゴリズムを用いていますが検索には時間がかかります。BLAST は生物学的な意味で厳密さに欠ける方法ですが、他に比べて高速なので広く実用的に使われています。FastA はデータベースの遺伝子との相同性が低い配列でも検索できると言われています。

遺伝子系統解析 Web システムの FastA での処理時間はサーバの状態にもよりますが5分～20分程度かかります。システムのメイン画面で「結果待ちリスト」ボタンを押すと現在実行中の処理を確認できます。他の人が使用中の場合は通常よりも時間がかかるのでお待ちください(現在デモ版なのでサーバ処理速度は遅いです)。

ここでは、「検索アルゴリズム」は FastA を選択して「入力内容の送信」ボタンを押してください。

・検索配列データ: テキストの貼り付け、ファイルから読込 (塩基配列のテキストファイル、またはFastA形式ファイル)

testdata.txt ファイル読込

```

aacacatgcaagtcgaacgggtgacgaggagcttgctcctccgatcagtgccgaacgggtgagtaacacgtgagtaacctgccccag
tccggaaattattggcgtaaaagagcttgtaggcggtttgtgcgctctgctgtgaaaaatccggggctcaaccccgacttgagtg
ggtcggatatacagtggtgcatggttgtcgtcagctcgtgctgtagatgttggttaagtcccgaacgagcgcaacctcgtctc

```

・検索配列名: (英数半角記号)

・検索対象データベース:

バクテリア 16SrDNA (DDB) DDBJ全てデータ(DDBJ)定期リリース+新着データ(※未対応)

・検索アルゴリズム:

FastA (処理時間 5 ~ 20分) Search (Smith-Waterman) (処理時間 1 ~ 3時間)

入力内容の送信 処理待ちリスト 処理結果参照

図 3-2 検索条件入力

実行が終了すると処理結果が表示されます。処理結果(図 3-3)では *Tetrasphaera elongata* の相関性が高いのでサンプルの細菌が *Tetrasphaera elongata* であると推定されます。

処理結果 DNA配列検索Webシステム Demo ver. 0.1 ×

testonda1 ファイル保存

サンプル名	DNA配列		
sample	aacacatgcaagtcgaacgggtgacgaggagcttgctcctccgatcagtgccgaacgggtgagtaacacgtgagtaacctgccccag		
アクセッションNo.	生物種(株)	相関性(%)	DNA配列
AB051430_1	<i>Tetrasphaera elongata</i>	100.0	aacacatgcaagtcgaacgggtgacgaggagc
AB030911_1	actinomycete Lp2 16S	99.6	gacgaacgctggcggtgcttaacacatgca
AB072496_1	<i>Tetrasphaera duodeca</i>	97.4	agagttgatcctggctcaggacgaacgctgg
EU707564_1	<i>Tetrasphaera</i> sp. YC67	97.5	catgcaagtcgaacgggtgaccgaggaagctt
Y14597_1	<i>Tetrasphaera jenkinsii</i>	97.1	gatttgatcctggctcaggacgaacgctggcg
X85212_1	<i>Tetrasphaera jenkinsii</i>	97.0	agtttgatcctggctcaggacgaacgctggcg
X85211_1	<i>Tetrasphaera jenkinsii</i>	97.0	agtttgatcctggctcaggacgaacgctggcg
AF125091_1	<i>Tetrasphaera australie</i>	96.8	ggttcaggacgaacgctggcggtgcttaac

キャンセル 全データ詳細 相関性の詳細 情報の取得 マルチプルアライメントへ

図 3-3 処理結果

仮に、実行中に他の処理を行う場合は実行待ち画面で「非同期実行」ボタンを押してメイン画面に戻ってください。(ブラウザを閉たり、パソコンを終了させた場合は再起動ログイン後に「処理結果参照」ボタンを押すと実行結果を確認できます。)

結果の一覧で「情報の取得」ボタンを押すと、カーソル位置の遺伝子の詳細情報を DDBJ のデータベースから獲得できます。前述のフラットファイル形式で表示されます(図 3-4)。

The image displays two screenshots of a web browser window showing detailed genetic information for two different genes. The top screenshot shows details for locus AB051430, and the bottom screenshot shows details for locus AB030911. Both screens display fields like LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS, SOURCE, ORGANISM, REFERENCE, AUTHORS, TITLE, JOURNAL, COMMENT, and FEATURES.

Top Screenshot (Locus AB051430):

```

LOCUS      AB051430          1390 bp  DNA   linear  BCT 05-DEC-2008
DEFINITION Tetrasphaera elongata gene for 16S rRNA, strain:ASP12.
ACCESSION  AB051430
VERSION    AB051430.1
KEYWORDS   .
SOURCE     Tetrasphaera elongata
ORGANISM   Tetrasphaera elongata
           Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
           Micrococcineae; Intrasporangiaceae; Tetrasphaera.
REFERENCE  1 (bases 1 to 1390)
AUTHORS    Onda,S. and Takii,S.
TITLE      Direct Submission
JOURNAL     Submitted (21-NOV-2000) to the DDBJ/EMBL/GenBank databases.
           Contact:Shin Onda
           Tokyo Metropolitan University, Dept. of Biology; 1-1 Minami-Osawa,
           Hachioji, Tokyo 192-0397, Japan
           URL      :http://www.metro-u.ac.jp/
REFERENCE  2
AUTHORS    Onda,S. and Takii,S.
TITLE      Isolation and characterization of a Gram-positive
           polyphosphate-accumulating bacterium
JOURNAL     J. Gen. Appl. Microbiol. 48, 125-133 (2002)
COMMENT
FEATURES   Location/Qualifiers
           source          1..1390
                       /db_xref="taxon:101689"
                       /mol_type="genomic DNA"
                       /organism="Tetrasphaera elongata"
    
```

Bottom Screenshot (Locus AB030911):

```

LOCUS      AB030911          1443 bp  DNA   linear  BCT 25-AUG-1999
DEFINITION Actinomycete Lp2 gene for 16S rRNA, partial sequence.
ACCESSION  AB030911
VERSION    AB030911.1
KEYWORDS   16S ribosomal RNA.
SOURCE     actinomycete Lp2
ORGANISM   Tetrasphaera elongata
           Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
           Micrococcineae; Intrasporangiaceae; Tetrasphaera.
REFERENCE  1 (bases 1 to 1443)
AUTHORS    Shintani,T., Hanada,S. and Nakamura,K.
TITLE      Direct Submission
JOURNAL     Submitted (09-AUG-1999) to the DDBJ/EMBL/GenBank databases.
           Contact:Tomoyoshi Shintani
           Industrial Research Center of Ehime Prefecture, Laboratory of Food
           Process; 487-2 Kumekubota, Matsuyama, Ehime 791-1101, Japan
           URL      :www.iri.pref.ehime.jp
REFERENCE  2
AUTHORS    Shintani,T., Hanada,S. and Nakamura,K.
TITLE      Actinomycete Lp2 gene for 16S rRNA, partial sequence
JOURNAL     Published Only in Database(1999)
COMMENT
FEATURES   Location/Qualifiers
           source          1..1443
                       /db_xref="taxon:101689"
                       /mol_type="genomic DNA"
                       /note="gram-positive high GC bacterium"
    
```

図 3-4 遺伝子情報詳細

4 複数の遺伝子の比較

同じ機能を持つ遺伝子でも塩基配列には突然変異による違いがみられ、塩基配列を比較すると違いの多い部分と少ない部分がみられます。機能的に重要な部分は突然変異により機能を失うことが多いので、重要な部分には違いが少ないと考えられています。よって、未知の遺伝子の機能を推定するにきは複数の配列の共通のパターンを解析すれば良いと考えられます。また、配列の変異から遺伝子の進化と系統を解析することもできます。複数の配列を整列して比較して遺伝子の解析を行うことをマルチプルアライメント(多重整列)といいます。

遺伝子系統解析 Web システムを使ってマルチプルアライメントを行ってみましょう。例として、前述の遺伝子相同性検索の結果として獲得された細菌の 16SrDNA 遺伝子を利用します。

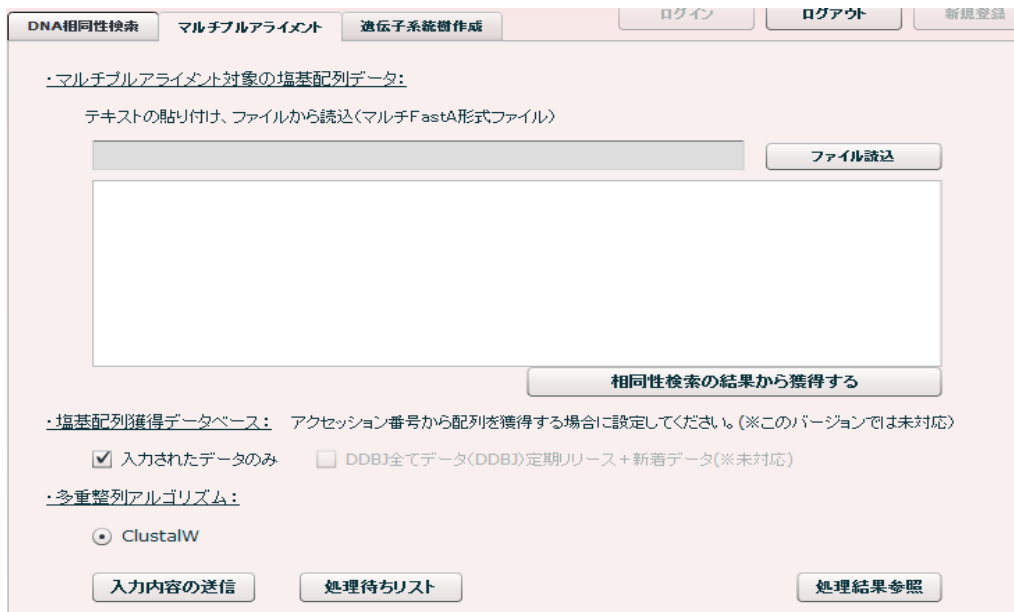


図 4-1 マルチプルアライメントタブ画面

まず、「マルチプルアライメント」タブを開きます(図 4-1)。

「相同性検索の結果から獲得する」ボタンを押して、結果リストを表示してください(図 4-2)。以前に行った相同性検索の日時のレコードを選択して「結果表示」ボタンを押すと前回の結果画面が表示されます。この画面で「マルチプルアライメントへ」のボタンを押すと、マルチプルアライメント画面へデータがコピーされます。

(デモバージョンでは、処理の都合で配列名は30文字以下、括弧や空白は _ に変換されます。)



名称	開始日付	状態	処理選択
testonda1	10/11/17 05::	END	SIM:FASTA
testonda1	10/09/28 19::	END	SIM:FASTA
testonda1	10/11/13 04::	END	SIM:FASTA
testonda1	10/09/28 19::	END	SIM:FASTA
testonda1	10/11/17 06::	END	SIM:FASTA
testonda1	10/11/17 06::	END	SIM:FASTA
testonda1	10/11/18 20::	END	SIM:FASTA

図 4-2 結果リスト画面

マルチプルアライメント画面で送信するデータは、図 4-3 に示すように、それぞれの遺伝子の行の先頭が半角の > で始まって配列名、塩基配列の順に並ぶ形式とします。この形式はマルチ FastA 形式と呼ばれています。（「ファイル読込」ボタンで、他から獲得・編集したマルチ FastA 形式のファイルを読み込むこともできます。）

・マルチプルアライメント対象の塩基配列データ:

テキストの貼り付け、ファイルから読込(マルチFastA形式ファイル)

ファイル読込

```
>sample
aacacatgcaagtgcgaacgggtgacgaggagcttgctcctccgatcagtggcgaacgggtg
agtaacacggtgagtaacctgcccagactctggaataacagttggaacagctgctaata
ccggatacggagacggagagggcatctctaccgtctggaagttttcggctctgggatggac
tcgcggtctatcagcttgttggtgaggtagtggtcaccgaaggcgacgacgggtagccgg
cctgagagggcgaccggccacactgggactgagacacggcccagactcctacgggaggca
gcagtggggaatattgcacaatgggcaaacgctgatgcagcaacgcccggtaggggatg
acggccttcgggttgtaaacctctttcagtagggaagaagcgaaagtgcaggtacctaca
```

相同性検索の結果から獲得する

・塩基配列獲得データベース: アクセション番号から配列を獲得する場合に設定してください。(※このバージョンでは未対応)

入力されたデータのみ DDBJ全てデータ(DDBJ)定期リリース+新着データ(※未対応)

・多重整列アルゴリズム:

ClustalW

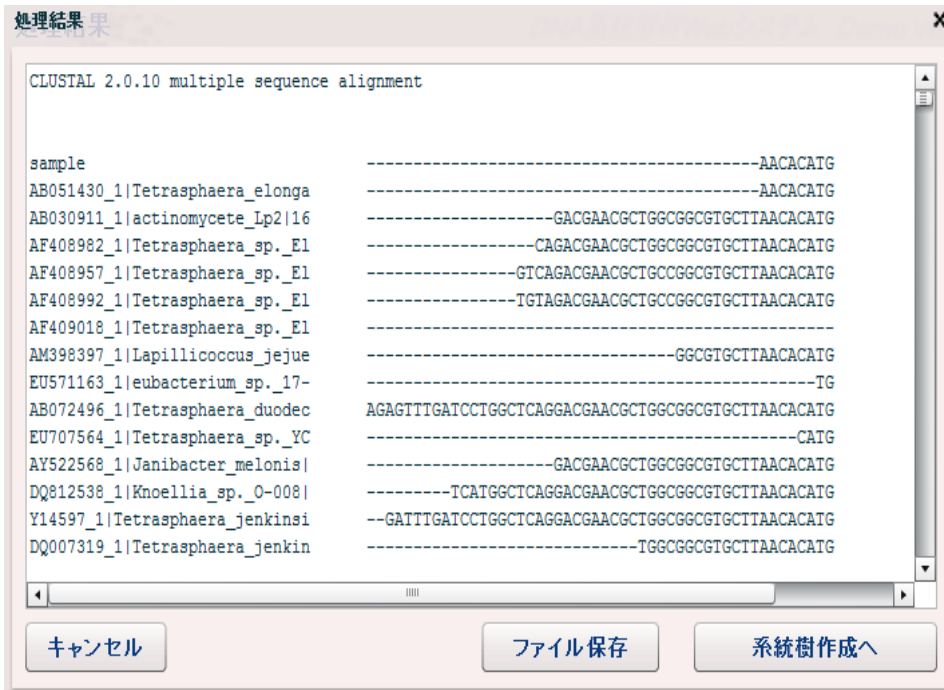
マルチ FastA 形式 例)

```
>sampleNo. 1
ATGCATGCATGCATGC
ATGCATGCATGCATGC

>sampleNo. 2
ATGCATGCATGCATGC
ATGCATGCATGCATGC
```

図 4-3 マルチプルアライメント検索入力画面 マルチ FastA 形式のデータ

「入力内容の送信」ボタンを押すと処理待ち画面が表示され、処理が終わると図 4-4 のような結果画面が現れます。結果の下の * の記号が変異のない部分(共通する部分)を表しています。記号の - はギャップで、この部分にフレームシフト突然変異が起こった可能性を示しています(先頭と最後の - については塩基配列が登録されていないという意味です)。



```

sample
AB051430_1|Tetrasphaera_elonga
AB030911_1|actinomycete_Lp2|16
AF408982_1|Tetrasphaera_sp._El
AF408957_1|Tetrasphaera_sp._El
AF408992_1|Tetrasphaera_sp._El
AF409018_1|Tetrasphaera_sp._El
AM398397_1|Lapillicoccus_jejue
EU571163_1|eubacterium_sp._17-
AB072496_1|Tetrasphaera_duodec
EU707564_1|Tetrasphaera_sp._YC
AY522568_1|Janibacter_melonis|
DQ812538_1|Knoellia_sp._0-008|
Y14597_1|Tetrasphaera_jenkinsi
DQ007319_1|Tetrasphaera_jenkin

```

```

-----AACACATG
-----AACACATG
-----GACGAACGCTGGCGCGTGCTTAACACATG
-----CAGACGAACGCTGGCGCGTGCTTAACACATG
-----GTCAGACGAACGCTGCCGCGTGCTTAACACATG
-----TGTAGACGAACGCTGCCGCGTGCTTAACACATG
-----
-----GGCGTGCTTAACACATG
-----TG
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGCGTGCTTAACACATG
-----CATG
-----GACGAACGCTGGCGCGTGCTTAACACATG
-----TCATGGCTCAGGACGAACGCTGGCGCGTGCTTAACACATG
--GATTTGATCCTGGCTCAGGACGAACGCTGGCGCGTGCTTAACACATG
-----TGCGGCGTGCTTAACACATG

```

```

sample
AB051430_1|Tetrasphaera_elonga
AB030911_1|actinomycete_Lp2|16
AF408982_1|Tetrasphaera_sp._El
AF408957_1|Tetrasphaera_sp._El
AF408992_1|Tetrasphaera_sp._El
AF409018_1|Tetrasphaera_sp._El
AM398397_1|Lapillicoccus_jejue
EU571163_1|eubacterium_sp._17-
AB072496_1|Tetrasphaera_duodec
EU707564_1|Tetrasphaera_sp._YC
AY522568_1|Janibacter_melonis|
DQ812538_1|Knoellia_sp._0-008|
Y14597_1|Tetrasphaera_jenkinsi
DQ007319_1|Tetrasphaera_jenkin
DQ007321_1|Tetrasphaera_jenkin
X85212_1|Tetrasphaera_jenkinsi
X85211_1|Tetrasphaera_jenkinsi
AF125091_1|Tetrasphaera_austra
AF125090_1|Tetrasphaera_austra
DQ007320_1|Tetrasphaera_vanvee

```

```

CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CCGATCAGTGGCGA
CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CCGATCAGTGGCGA
CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CCGATCAGTGGCGA
CAAGTCGAACGGTGACCTCGAGAGCTT-GCTCTTGGGTGATCAGTGGCGA
CAAGTCGAACGGTGACCTCGAGAGCTT-GCTCTCGGGTATCAGTGGCGA
CAAGTCGAACGGTGACCTCGAGAGCTT-GCTCTTGGGTGATCAGTGGCGA
-----CGCGAGAGCTTTGCTCTTGGGTGATCAGTGGCGA
CAAGTCGAACGGTGACCTCGAGAGCTT-GCTCTTGGGTGATCAGTGGCGA
CAAGTCGAACGGTGACGACAGGAGCTT-GCTCCGGTCTGATCAGTGGCGA
CAAGTCGAACGGTGAAGGTGGGAGCTT-GCTTCTACCGGATCAGTGGCGA
CAAGTCGAACGGTGACCGAGGAAGCTT-GCTCCT-CGTGATCAGTGGCGA
CAAGTCGAACGGTGAACCTTGGAGCTT-GCTCTAAGGGGATCAGTGGCGA
CAAGTCGAACGGTGATCTTGGGAGCTT-GCTCCTGGGTGAGCAGTGGCGA
CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CTGATCAGTGGCGA
CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CTGATCAGTGGCGA
CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CTGATCAGTGGCGA
CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CTGATCAGTGGCGA
CAATTCGAACGGTGACC--AGGAGCTT-GCTCCT--GTGATCAGTGGCGA
CAAGTCGAACGGTGACC--AGGAGCTT-GCTCCT--GTGATCAGTGGCGA
CAAGTCGAACGGTGACG--AGGAGCTT-GCTCCT--CTGATCAGTGGCGA

```

```

***** ***          ** *****

```

図 4-4 マルチプルアライメント実行結果

5 遺伝子から進化の推定

異なる生物で共通な、同じ働きをする遺伝子のDNAやRNAの塩基配列を比較することによって、遺伝子の進化的な近縁関係を推定することができますとされています。この遺伝子レベルでの近縁関係を求めることを遺伝子系統解析、それを図にしたものを遺伝子系統樹(図 5-1、図 5-2)と呼びます。系統樹の表記方法にも有根系統樹と無根系統樹があり、一般的には進化の起源を示さずに遺伝的な類似性をのみを表す場合は無根系統樹を使います(図 5-2)。



図 5-1 遺伝子系統樹の例(PPAR- γ 遺伝子の系統)

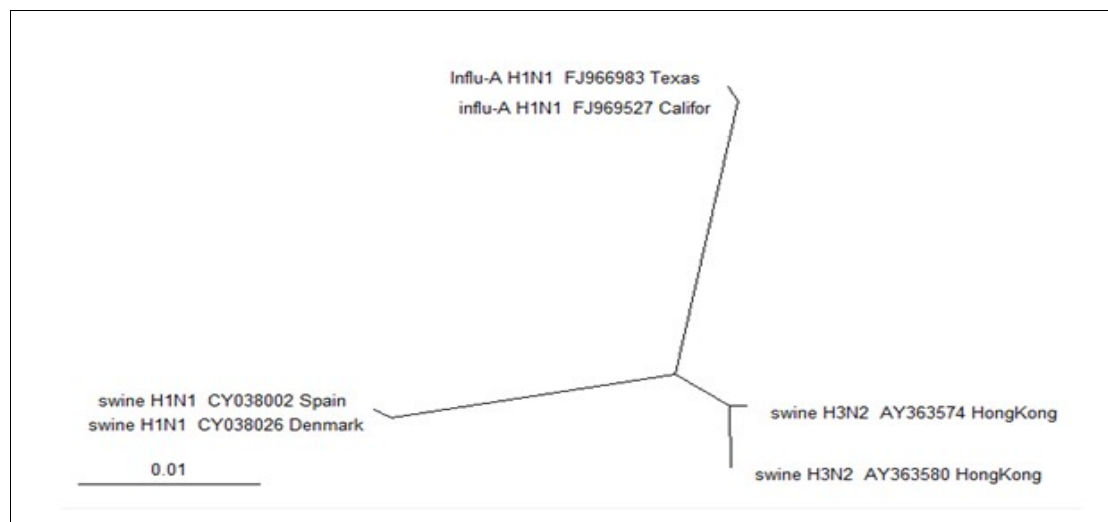


図 5-2 遺伝子系統樹の例(インフルエンザの膜タンパクの系統)

遺伝子系統解析の手順は、まず対象となる遺伝子のマルチプルアライメントを行い、それから進化距離(遺伝子の類似の程度)を求め遺伝子系統樹を作成します。遺伝子系統樹を作成するアルゴリズムには、近接結合法(Neighbor-Joining Method/ NJ法と略す)、最大節約法(maximum parsimony method)など、いくつかの方法がありますが、それぞれ生物学的な意味合いの異なるもので目的や背景によって合理的な方法を選びます。

前述のマルチプルアライメントの結果を利用して、系統樹を書いてみましょう。遺伝子系統解析 Web システムで「遺伝子系統樹作成」タブを開き「マルチプルアライメントの結果から獲得する」ボタンを押して前回のマルチプルアライメントの結果を獲得してください。結果リスト画面から前回の処理日時レコードを選択して処理結果画面を開き「系統樹作成へ」ボタンを押すと獲得できます(図 5-3)。

「遺伝子系統樹作成」タブ画面で「入力内容の送信」ボタンを押してしばらく待つと図 5-4 のような処理結果画面が表示されます。



図 5-3 遺伝子系統樹作成タブ画面



図 5-4 遺伝子系統解析の処理結果(Newick 形式)

この結果(図 5-4)の表しているものは、系統関係と進化距離をテキスト形式で表したもので Newick 形式と呼ばれています。画面の「系統樹描画編集ツール起動」を押すと系統樹描画編集 Web ページが別のウインドウ(またはタブ)で開き遺伝子系統樹が現れます(図 5-5)。

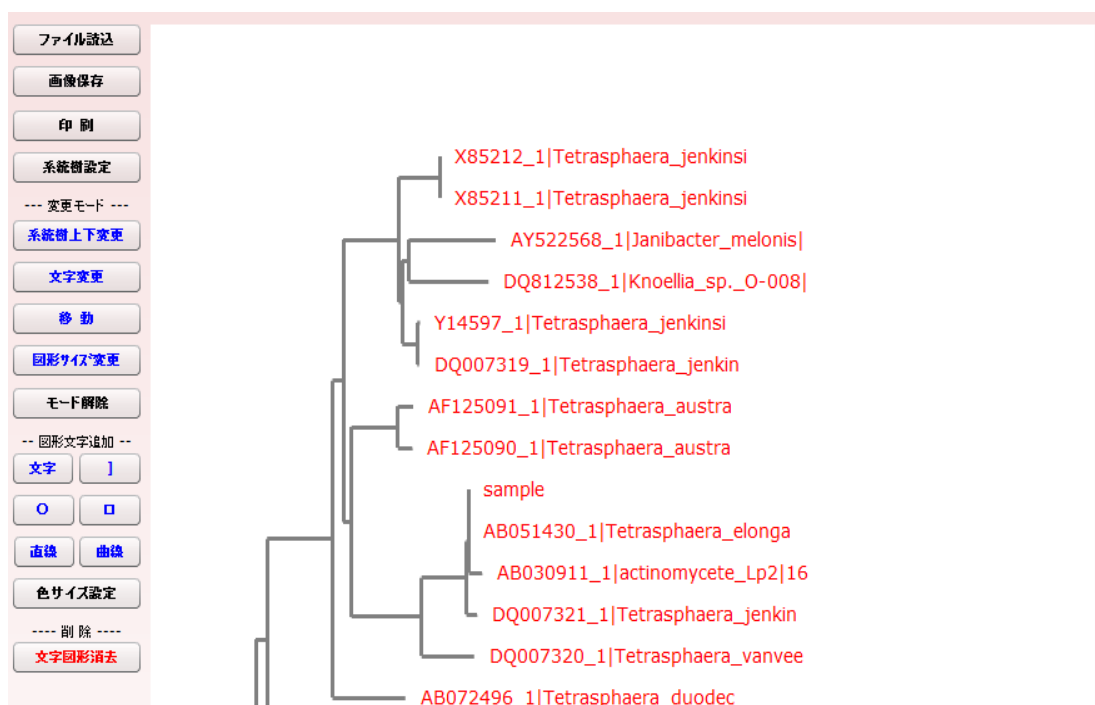


図 5-5 系統樹描画編集 Web ページと処理結果の遺伝子系統樹

この結果(図 5-5)の場合は、sample の細菌は、Tetrasphaera 属の系統に属していることがわかります。

この例の場合はサンプルとして既知の塩基配列を用いましたが、未知の微生物から抽出した塩基配列で相同性検索や遺伝子系統解析を行うことによって、未知の細菌の系統関係を求めることができます。近縁のデータがない場合は新種・新型の微生物である可能性が考えられます。十分に研究されている分類群の場合はその細菌の種類が判別できるので、遺伝子を利用した微生物の検出技術に応用することができます。

===== END =====