７．タンパク質モチーフ検索システム (MOTIF) の利用法

目次

はじめに *7-2*
講習内容 *7-2*
１．検索依頼電子メールの作成 *7-2*
２．検索結果の受取と加工 *7-4*
３．結果ファイルの解凍 *7-8*
４．グラフィックスの表示法 *7-9*

はじめに

タンパク質配列モチーフ検索システム MotifFinder は、与えられた問い合わせ配列の中にモチーフライブラリに登録されているような配列モチーフがあるかどうかを検索し、見つかった場合にはその種類と場所と共に、そのモチーフを共有するほかのタンパク質でそのモチーフ配列と他の機能部位などの分布がどうなっているのかを図式的に表示するためのコード、さらにはこのようなモチーフを共有するタンパク質で構造既知のものがあった場合は、その 3 次元構造のステレオ図を表示するためのグラフィックコードを返すシステムである。これは従来のモチーフ探索プログラムと比較して、単にその存在と場所だけでなく、関連するタンパク質でのアノテーション情報も得られるために、タンパク質配列の意味づけを行おうとする研究者の助けとなることが期待される。なお、このシステムは現在電子メールによるメールサーバーとして提供されている。

講習内容

検索依頼電子メールの作成
検索結果の受取と加工
結果ファイルの解凍
グラフィックスの表示法

1．検索依頼電子メールの作成

MotifFinder システムはメールサーバーとしてのみ提供されているので、まず問い合わせ配列を電子メール上で作成・編集する必要がある。

なお、電子メールの使用法そのものについては別の講習があるので、ここでは詳しくは述べない。また電子メールを利用する手段も幾通りかあり、ここでの説明の便宜上必ずしもメール使用法講習のものと同じでないかも知れないが、もちろん利用者が好みの方法で利用されて構わない。

まず最初にこのシステムの利用法を書いた文書を取り寄せるためのメールを出してみよう。他のメールサーバーと同様、このシステムも、本文に `help` とだけ書いたものを出すと使用法説明が得られる。

MotifFinder システムのメールアドレスは

```
motif@genome.ad.jp
```

である。

使用法の請求：

```
hgc{ogi}96:mail motif@genome.ad.jp
Subject:        (改行のみ)
help            (改行の後、CTRL+D を押す)
Cc:             (改行のみ)
hgc{ogi}97:
```

通常 HGC ではメールシステムとして MH が用いられるが、このように編集のいらない極めて簡単なものは直接 mail コマンドを呼び出すという手もある。mail コマンドは、続けて宛先アドレスを付けて呼び出す。まず Subject: を聞いてきたら、これは必要ないのでリターンのみでよい。次の行に直接 help と書き、リターンを打つ。続けてコントロールキーを押しながら D のキーを押すと、(Cc: を聞いた後) メールが発信される。(もし Cc: と聞いてきたら、これも不要なのでリターン。)

検索依頼メールの形式：

検索依頼メールには、問い合わせたいタンパク質配列を FASTA フォーマットで記述するだけでよい。即ち、第 1 行目は > ではじめ、ここに配列の名前とコメントを書く。 2 行目以下に配列をアミノ酸 1 文字コードで書くだけである。ここでの空白及び改行は自由にいれてよい。

この講習では問い合わせ配列として SWISS-PROT から AACT_HUMAN (ヒトのαアクチニン) というエントリーを取ってきて使用することとする。 bget の -f オプションを使用して、直接 FASTA 形式で配列を query.seq と言う名前のファイルに格納しよう。

```
bget swissprot -f aact_human >query.seq
```

このファイルの中身を確認してみる。

```
hgc{ogi}110: cat  query.seq
>AACT_HUMAN (swissprot)
MDHYDSQQTNDYMQPEEDWDRDLLLDPAWEKQQRKTFTAWCNSHLRKAGTQIENIEEDFR
DGLKLMLLLEVISGERLAKPERGKMRVHKISNVNKALDFIASKGVKLVSIGAEEIVDGNV
KMTLGMIWTIILRFAIQDISVEETSAKEGLLLWCQRKTAPYKNVNIQNFHISWKDGLGFC
ALIHRHRPELIDYGKLRKDDPLTNLNTAFDVAEKYLDIPKMLDAEDIVGTARPDEKAIMT
YVSSFYHAFSGAQKAETAANRICKVLAVNQENEQLMEDYEKLASDLLEWIRRTIPWLENR
VPENTMHAMQQKLEDFRDYRRLHKPPKVQEKCQLEINFNTLQTKLRLSNRPAFMPSEGRM
VSDINNAWGCLEQVEKGYEEWLLNEIRRLERLDHLAEKFRQKASIHEAWTDGKEAMLRQK
DYETATLSEIKALLKKHEAFESDLAAHQDRVEQIAAIAQELNELDYYDSPSVNARCQKIC
DQWDNLGALTQKRREALERTEKLLETIDQLYLEYAKRAAPFNNWMEGAMEDLQDTFIVHT
IEEIQGLTTAHEQFKATLPDADKERLAILGIHNEVSKIVQTYHVNMAGTNPYTTITPQEI
NGKWDHVRQLVPRRDQALTEEHARQQHNESVRKQFGAQANVIGPWIQTKMEEIGRISIEM
HGTLEDQLSHLRQYEKSIVNYKPKIDQLEGDHQLIQEALIFDNKHTNYTMEHIRVGWEQL
LTTIARTINEVENQILTRDAKGISQEQMNEFRASFNHFDRDHSGTLGPEEFKACLISLGY
DIGNDPQGEAEFARIMSIVDPNRLGVVTFQAFIDFMSRETADTDTADQVMASFKILAGDK
NYITMDELRRELPPDQAEYCIARMAPYTGPDSVPGALDYMSFSTALYGESDL
```

これはもはや編集する必要がないので、直接 MotifFinder にメールで送ってみよう。

```
mail motif@genome.ad.jp <query.seq
```

なお、検索対象モチーフライブラリはこのまま使用すると PROSITE が使われるが、このほかに京大化研で作られている、類似タンパク質ファミリーを特徴づけるような保存配列フラグメントから作られた MotifDic も使用可能である。この場合は問い合わせ配列の前に datalib MotifDic と書いた行をいれること。

2．検索結果の受取と加工

メールを送信し終わってから戻るまでしばらくかかるが、例えばリターンキーを叩くと、新着メールがあると

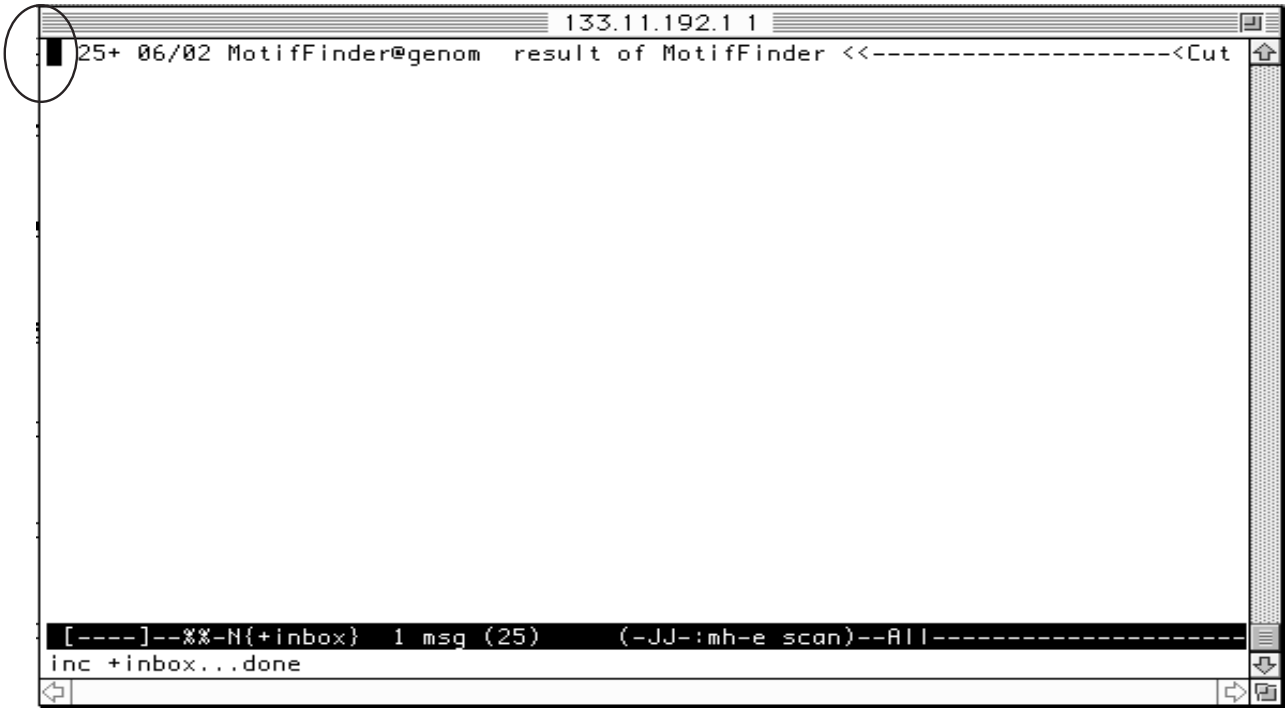
```
hgc {ogi} 115:
You have new mail.
```

のように表示されるのでメールが来たことが分かる。

メールの読み込みと加工は、ここでは Nemacs エディターの MH-rmail を使って行ってみる。まず、

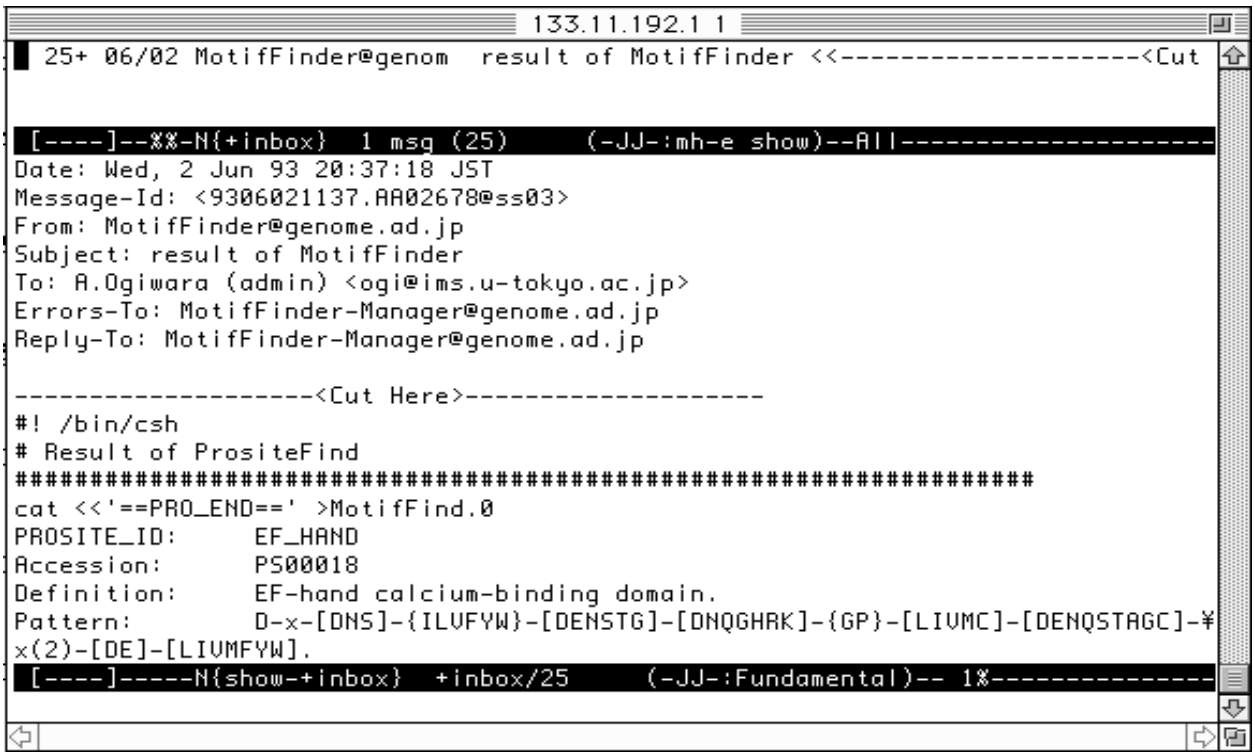
```
emacs -nw -f mh-rmail
```

と入力する。すると画面が以下のように変わる。

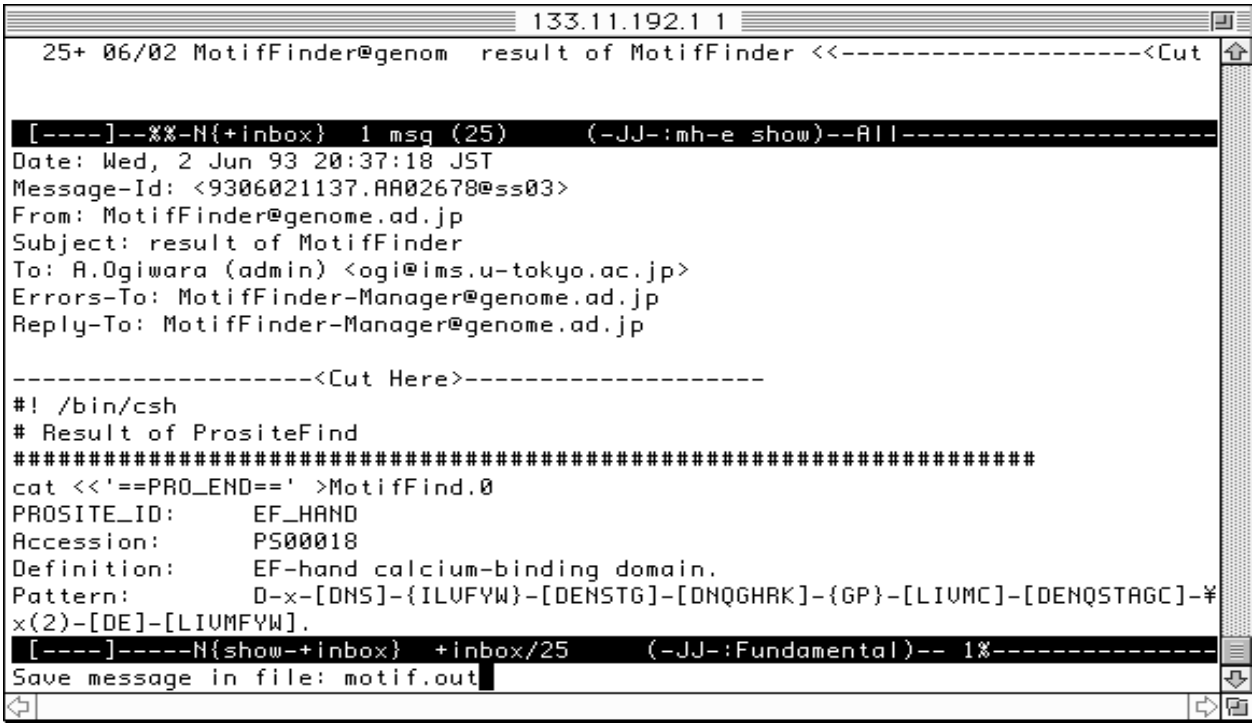


今まで未読のメールがないならば、上の画面のように MotifFinder からの返事 1 通のみが表示され、そのリストの頭にカーソルが来ている。（もしカーソルが別の位置にあるならば、n または p を押して前後に移動して MotifFinder からの返信にカーソルを合わせる。）

ここでピリオド（.）を押すと、メールの内容が次のように表示される。

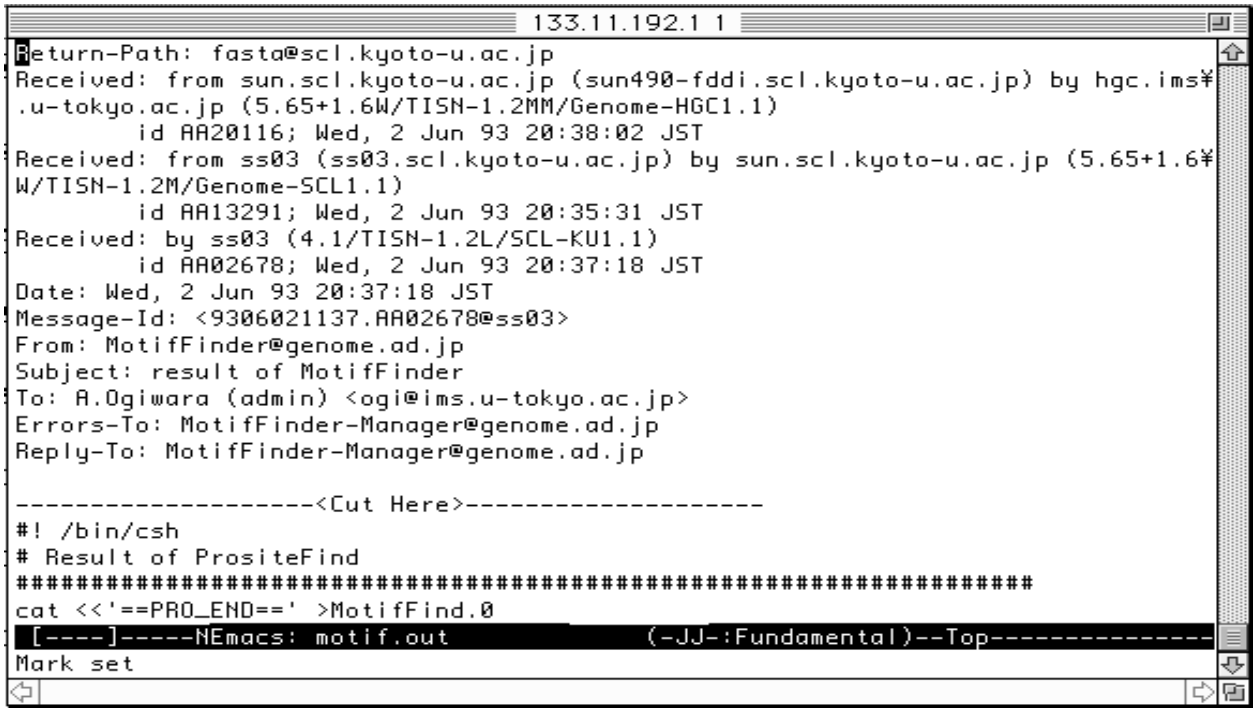


これを一旦ファイルに保存しよう。カーソルがこのままの位置で > を打つとセーブするファイル名を聞いてくる。ここでは `motif.out` と言う名前にしよう。



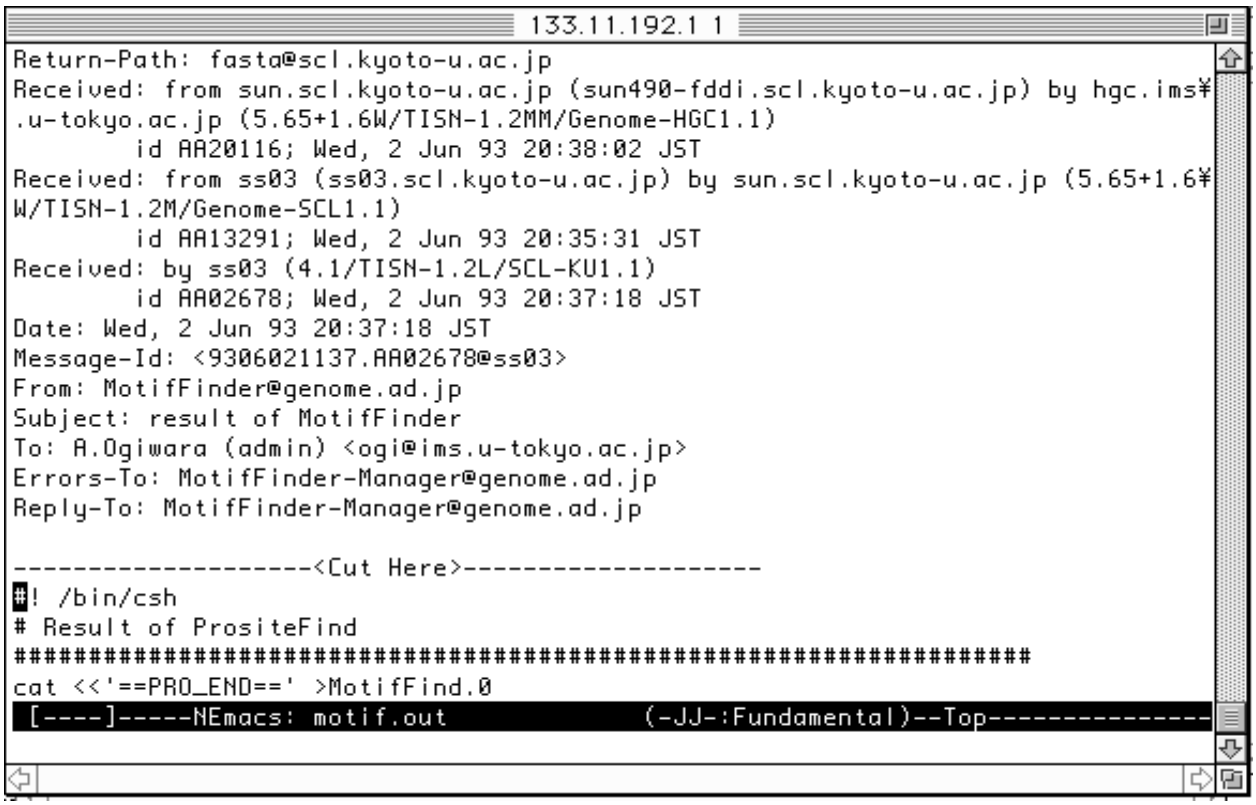
ファイル名を指定してリターンを押したならば保存は完了するので、`q` キーを押して mh-rmail モードを終了する。続いて今セーブしたファイルを編集するために `C-x C-f` コマンドを入力する。なお emacs の説明では、C-x はコントロールキーを押しながら x キーを押すことを意味し、M-x はエスケープキーを押した後に x キーを押すことを意味する。つまり、ここではコントロールキーを押しながら x キーを押し、続けてコントロールキーを押しながら f キーを押す。

編集したいファイル名を問い合わせてくるので、今セーブしたmotif.out を入力する。

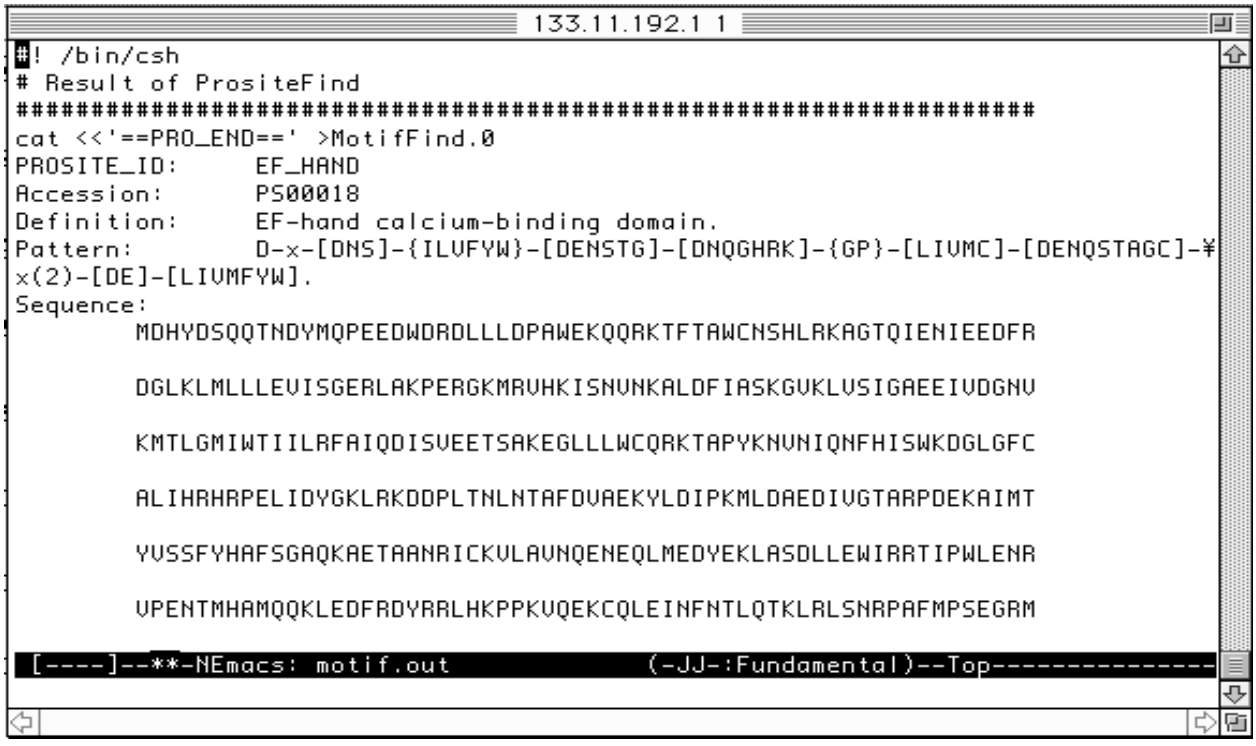


ここで行う必要のあることは、メールのヘッダー部分から---<Cut Here>--- と書かれた行までを消去することである。今はファイル先頭にカーソルがあるので、先ずはC-ｽﾍﾟｰｽを押してその位置をマークする（上の図）。

次いでC-nを何回か押してカーソルを---<Cut Here>--- の次の行の先頭に位置づける。



この位置で C-w を押すと、先ほどマークしたところから今カーソルがある場所までが消去される。



これで必要な編集作業は終了である。C-x C-s を押して編集結果をセーブし、次いで C-x C-c を押して emacs を終了させる。

3．結果ファイルの解凍

今加工してセーブしたファイル motif.out には、MotifFinder の結果がいくつかのファイルに分かれたものが 1 まとまりになって入っている。そこでこれを解凍してばらばらにしてやる必要がある。そのためには、今の結果ファイルに csh を適用する。

解凍前のファイルはこれだけあったが、

```
hgc{ogi}126: ls
motif.out   motif.out~  query.seq
```

ここで `csh motif.out` を実行すると、

```
hgc{ogi}127: csh motif.out
```

この後ファイルリストを見ると、以下のようにたくさんのファイルができている。

```
hgc{ogi}128: ls
Motif3D0.ps    MotifFeat1.ps  MotifFind.0    MotifFind.2    motif.out~
MotifFeat0.ps  MotifFeat2.ps  MotifFind.1    motif.out      query.seq
hgc{ogi}129:
```

なお、この時の注意として、すでに Motif...と言う同名のファイルが存在するとそれらを上書きしないようになっているため、あらかじめこういった名前のファイルを消去しておくこと。

このうち MotifFind.番号 となっているファイルは発見されたモチーフの名前とその場所が問い合わせ配列の上に示されている。番号は複数のモチーフが見つかったときにそれらを識別するための通し番号である。

```
hgc{ogi}129: cat MotifFind.0
PROSITE_ID:     EF_HAND
Accession:      PS00018
Definition:     EF-hand calcium-binding domain.
Pattern:        D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x
(2)-[DE]-[LIVMFYW].
Sequence:
        MDHYDSQQTNDYMQPEEDWDRDLLLDPAWEKQQRKTFTAWCNSHLRKAGTQIENIEEDFR

        DGLKLMLLLEVISGERLAKPERGKMRVHKISNVNKALDFIASKGVKLVSIGAEEIVDGNV

        KMTLGMIWTIILRFAIQDISVEETSAKEGLLLWCQRKTAPYKNVNIQNFHISWKDGLGFC

        ALIHRHRPELIDYGKLRKDDPLTNLNTAFDVAEKYLDIPKMLDAEDIVGTARPDEKAIMT

        YVSSFYHAFSGAQKAETAANRICKVLAVNQENEQLMEDYEKLASDLLEWIRRTIPWLENR

        VPENTMHAMQQKLEDFRDYRRLHKPPKVQEKCQLEINFNTLQTKLRLSNRPAFMPSEGRM

        VSDINNAWGCLEQVEKGYEEWLLNEIRRLERLDHLAEKFRQKASIHEAWTDGKEAMLRQK

        DYETATLSEIKALLKKHEAFESDLAAHQDRVEQIAAIAQELNELDYYDSPSVNARCQKIC

        DQWDNLGALTQKRREALERTEKLLETIDQLYLEYAKRAAPFNNWMEGAMEDLQDTFIVHT

        IEEIQGLTTAHEQFKATLPDADKERLAILGIHNEVSKIVQTYHVNMAGTNPYTTITPQEI

        NGKWDHVRQLVPRRDQALTEEHARQQHNESVRKQFGAQANVIGPWIQTKMEEIGRISIEM

        HGTLEDQLSHLRQYEKSIVNYKPKIDQLEGDHQLIQEALIFDNKHTNYTMEHIRVGWEQL

        LTTIARTINEVENQILTRDAKGISQEQMNEFRASFNHFDRDHSGTLGPEEFKACLISLGY
                                                ++++++++++++
        DIGNDPQGEAEFARIMSIVDPNRLGVVTFQAFIDFMSRETADTDTADQVMASFKILAGDK

        NYITMDELRRELPPDQAEYCIARMAPYTGPDSVPGALDYMSFSTALYGESD
```