Last modified August 13, 2018

Elementary Sequence Analysis

Brian Golding, Dick Morton and Wilfried Haerty

Department of Biology
McMaster University
Hamilton, Ontario
L8S 4K1

These notes are in Adobe Acrobat format (they are available upon request in other formats) and they can be obtained from the website http://helix.biology.mcmaster.ca/courses.html. Some of the programs that you will be using in this course and which will be run locally can be found at http://evol.mcmaster.ca/p3S03.html.

The "blue text" should designate links within this document while the "red text" designate links outside of this document. Clicking on the latter should activate your web browser and load the appropriate page into your browser. If these do not work please check your Acrobat reader setup. The web links are accurate to the best of our knowledge but the web changes quickly and we cannot guarantee that they are still accurate. The links designated next to the JAVA logo,  , require that JAVA be installed on your computer.

These notes are used in Biology 3S03. The purpose of this course is to introduce students to the basics of bioinformatics and to give them the opportunity to learn to manipulate and analyze DNA/protein sequences. Of necessity only some of the more simple algorithms will be examined.

The course will hopefully cover . . .

- databases of relevance to molecular biology.
- some common network servers/sites that provide access to these databases.
- methods to obtain sequence analysis software and data.
- methods of sequence alignment.
- methods of calculating genetic distance.
- methods of phylogenetic reconstruction.
- methods for detecting patterns and codon usage.
- methods for detecting gene coding regions.

The formal part of the course will consist of two approximately one hour lectures each week. Weekly assignments will be be provided to practice and explore the lecture material. In addition there will be an optional tutorial to help students with these assignments or other problems. These assignments will be 40% of your grade and three, in class quizzes will make up the remainder.

We would appreciate any comments, corrections or updates regarding these notes.

Golding@McMaster.CA Morton@McMaster.CA HaertyW@McMaster.CA

This document was entirely constructed with pdfLaTeX. Enormous kudos to all those that make this great software free to everyone.

Table of Contents in Brief

In order to speed download, I place here links to the individual chapters in pdf format. The contents of these are shown on the following 'Contents' pages but note that the links will function only for the individual chapter included here.

Contents

Chapter 8

Database Searching

8.1 Are there homologues in the database?

The following are some of the common programs currently being used to search the databases to find sequences similar to a specific query sequence provided by the user. In addition to finding out the identity of an unknown sequence they are also useful to find homologues and ancestral sequences that have similar or related functions/sequences.

8.1.1 FASTA

To search through the whole genetic sequence database can take a great deal of time due to its enormous size. If some operation must be performed on each sequence in turn then this can take even longer. One such example is to look throughout the whole database for homologous or similar sequences. To do this, special programs have been developed to speed the search. The first amongst these programs was a program called FASTA written by W.R. Pearson and D.J. Lipman (1988, PNAS 85:2444-2448).

It is possible to run this program on remote machines. The obvious choice for such a remote machine would be one that has access to the latest sequence information. Both EMBL and DDBJ have permitted this type of access and have implemented FASTA type searches through their machines (NCBI prefers to use BLAST - see below).

There are several flavours to FASTA: `fasta` scans a protein or DNA sequence library for sequences similar to a query sequence. `tfasta` compares a protein query sequence to a translated DNA sequence library. `lfasta` compares two query sequences for local similarity between them and shows the local sequence alignments. `plfasta` compares two sequences for local similarity and plots the local sequence alignments. Two recent flavours `fastx` and `fasty` (Pearson *et al.* 1997 Genomics 46:24-36) permit comparison of a DNA sequence translated in all six frames to the protein databases. The 'x' form takes a DNA query sequence and translates it in all frames and then permits gaps between the resulting amino acids. The 'y' form more generally permits gaps within and between codons. The related `tfastx` and `tfasty` forms compare a protein query sequence to a DNA database by translating the DNA database in all six frames.

I will illustrate what a FASTA type of search is and what the results look like with an example. Basically the idea is to search through the complete database for any possible similar sequence.

8.1.1.1 Instructions

To carry out this type of search on the EMBL server the following must be done. Either point your web browser to FASTA3 and fill out the appropriate forms or set up a file containing the following

```
LIB UNIPROT
WORD 1
LIST 50
TITLE HALHA
```

```
HISTOGRAM yes
SEQ
PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWP
FSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHL
DYLQNRVI
```

The first line contains the data library files to be searched (in this case all known protein entries). For protein searches this field may be one of

```
UniProt               A non-redundant collection of all proteins
UniRef100             As for UniProt but eliminate identical proteins
UniRef90              As for UniProt but eliminate proteins > 90% identical
UniRef50              As for UniProt but eliminate proteins > 50% identical
UniParc               As for UniProt but include archived proteins
                      (shows changes to an entry).
swiss-prot            Proteins in the SwissProt database
ipi                   Proteins in the International Protein Index
prints                Proteins in the FingerPrints database
sgt                   Proteins in the Structural Genomics Targets database
pdb                   Proteins in the 3D structural database PDB at Rutgers
imgthlap              Proteins in the Immunogenetics Database
Euro Patents          Proteins in the European patents database
Japan Patents         Proteins in the Japanese patents database
USPTO Patents         Proteins in the American patents database
```

For nucleotide searches this field may be one of

```
EMBL                  The entire EMBL database
FUNGI                 Subsection of EMBL.
INVERTEBRATES
HUMAN
MAMMALS
ORGANELLES
BACTERIOPHAGE
PLANT
PROKARYOTES
RODENTS
MOUSE
STSs                  Sequence Tagged Sites
SYNTHETIC
UNCLASSIFIED
VIRUSES
VERTEBRATES
ESTs                  Expressed Sequence Tags
GSSs                  Genome Survey Sequences
HTGs                  High throughput Genomics sequences
PATENTS
VECTORS
EMBLNEW               Sequences new since the last major database release
EMBLALL               EMBL + EMBLNEW
IMGTLIGM              Immunoglobulins and T cell receptors database
IMGTHLA               Human Major Histocompatiblity Complex (MHC/HLA) database
HGVBASE               Human Genome Variation database
```

The second line gives the word size or k-tuple value (more on this below). The third line says to LIST on the output the top 50 scores. The TITLE line is used for the subject of the mail message. Finally SEQ implies that everything below this line to the end of the message is part of the sequence. In this case the sequence is the protein sequence of the ferredoxin gene of *Halobacterium species NRC-1*.

The remaining options are - LIST n, n top scores listed in the output [50]. ALIGN n, align the top n to the query sequence [10]. ONE, compare only the given strand to the database, the default is to use the complementary strand as well. PROT will force your query sequence to be a protein (small protein sequences may be otherwise misinterpreted as DNA). PATH string mails the results back to string rather than the originator of the message.

After creating this file, mail the file by electronic mail to fasta@ebi.ac.uk and the results will be sent back to you by electronic mail. Alternatively simply point your web browser to FASTA3 and fill in the forms (they have the same options). Please, as a courtesy to others using the system please send only one job at a time. Many other people from all over the world are using these servers and the FASTA program is quite computer intensive despite its speed.

8.1.1.2 FASTA output

An example of the output is shown below. The input file is specifying the *Halobacterium species NRC-1* ferredoxin amino acid sequence to search the SWISS-PROT database.

```
FASTA searches a protein or DNA sequence data bank
 version 3.4t23 March 18, 2004
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query library @ vs +uniprot library
searching /ebi/services/idata/v916/fastadb/uniprot library

  1>>>HALHA - 128 aa
 vs  UniProt library

       opt      E()
< 20   1429      0:=
  22     30      0:=          one = represents 2608 library sequences
  24     50      1:*
  26    246     33:*
  28    917    356:*
  30   3937   2160:*=
  32  13990   8352:===*==
  34  33145  22650:========*====
  36  63533  46517:=================*=======
  38  94343  76875:==============================*=======
  40 124411 107234:=========================================*======
  42 156468 131081:=================================================*=========
  44 150019 144594:============================================================*==
  46 141925 147273:============================================================ *
  48 131289 140997:================================================    *
  50 124662 128660:=============================================== *
  52 106819 113114:==========================================  *
  54  83834  96619:==================================    *
  56  69228  80707:===========================   *
  58  58053  66259:=======================  *
  60  46296  53673:=================  *
  62  36284  43030:==============  *
  64  29298  34222:============ *
  66  22040  27048:========== *
  68  17474  21275:======= *
  70  13385  16672:======*
  72  10307  13028:=====*
  74   7966  10157:===*
  76   6188   7906:===*
  78   4631   6145:==*
  80   3596   4772:=*
  82   2939   3650:=*
  84   2320   2891:=*
  86   1609   2237:*
  88   1387   1731:*          inset = represents 13 library sequences
  90    920   1339:*
  92    642   1036:*         :=======================================*
  94    524    802:*         :========================================*
  96    393    620:*         :================================        *
  98    284    480:*         :======================              *
 100    257    371:*         :====================        *
 102    209    287:*         :=================     *
 104    124    222:*         :==========       *
 106     97    172:*         :========     *
 108     91    133:*         :=======   *
 110     83    103:*         :=======*
 112     46     80:*         :====  *
 114     39     62:*         :=== *
 116     38     48:*         :===*
 118     25     37:*         :==*
>120    604     29:*         :==*=======================================
501934690 residues in 1568424 sequences
 statistics extrapolated from 60000 to 1567839 sequences
  Expectation_n fit: rho(ln(x))= 5.0598+/-0.000193; mu= 9.7296+/- 0.011
 mean_var=58.3610+/-12.418, 0's: 151 Z-trim: 257  B-trim: 877 in 1/64
 Lambda= 0.167885
 Kolmogorov-Smirnov  statistic: 0.0646 (N=29) at  44

FASTA (3.47 Mar 2004) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  32
 Scan time: 272.383
The best scores are:                                      opt bits E(1568424)
```

```
UNIPROT:FER_HALN1 P00216 Ferredoxin.               ( 128)  870 217.3   1e-55
UNIPROT:Q9YGB6 Q9YGB6 Ferredoxin.                  ( 129)  761 190.9 9.1e-48
UNIPROT:FER_HALMA P00217 Ferredoxin.               ( 128)  750 188.2 5.8e-47
UNIPROT:FER_SYNP4 P15788 Ferredoxin.               (  98)  271  72.1 3.9e-12
UNIPROT:FER_SYNEL P00256 Ferredoxin I.             (  97)  263  70.2 1.5e-11
UNIPROT:FER_SYNLI P00255 Ferredoxin.               (  96)  262  69.9 1.7e-11
UNIPROT:FER_PHYPA O04166 Ferredoxin, chloroplast   ( 145)  254  68.1 9.3e-11
UNIPROT:FER1_ANASP P06543 Ferredoxin I.            (  98)  252  67.5 9.4e-11
UNIPROT:FER1_ANAVA P00254 Ferredoxin I.            (  98)  251  67.3 1.1e-10
UNIPROT:FER_NOSMU P00253 Ferredoxin.               (  98)  247  66.3 2.2e-10
UNIPROT:FER1_PLEBO Q51577 Ferredoxin I (FdI).      (  99)  245  65.8 3.1e-10
UNIPROT:FER3_CYACA P00241 Ferredoxin.              (  98)  242  65.1   5e-10
UNIPROT:Q7V0B6 Q7V0B6 Ferredoxin.                  (  99)  242  65.1 5.1e-10
UNIPROT:Q7VAM6 Q7VAM6 Ferredoxin.                  (  99)  241  64.8   6e-10
UNIPROT:Q7U8S7 Q7U8S7 Ferredoxin.                  (  99)  241  64.8   6e-10
UNIPROT:FER2_NOSMU P00249 Ferredoxin II.           (  98)  238  64.1 9.8e-10
UNIPROT:FER_CHLFR P00247 Ferredoxin.               (  98)  238  64.1 9.8e-10
UNIPROT:Q7M191 Q7M191 Ferredoxin.                  (  98)  238  64.1 9.8e-10
UNIPROT:FER1_CYAPA P17007 Ferredoxin I.            (  98)  237  63.9 1.2e-09
UNIPROT:FER1_NOSMU P00252 Ferredoxin I.            (  98)  236  63.6 1.4e-09
UNIPROT:FER_SYNY4 P00243 Ferredoxin.               (  96)  235  63.4 1.6e-09
UNIPROT:FER_EUGVI P22341 Ferredoxin.               (  96)  234  63.1 1.9e-09
UNIPROT:FER_SYNY3 P27320 Ferredoxin I.             (  96)  233  62.9 2.2e-09
UNIPROT:Q7TUS8 Q7TUS8 2Fe-2S Ferredoxin:Ferredoxi  (  99)  233  62.9 2.3e-09
UNIPROT:FER_MASLA P00248 Ferredoxin.               (  98)  232  62.7 2.7e-09
UNIPROT:FER1_SYNP7 P06517 Ferredoxin I.            (  98)  232  62.7 2.7e-09
UNIPROT:FER2_SPIOL P00224 Ferredoxin II.           (  97)  231  62.4 3.2e-09
UNIPROT:FER_CHLFU P56408 Ferredoxin.               (  94)  230  62.2 3.6e-09
UNIPROT:Q7M1S3 Q7M1S3 Ferredoxin C.                (  96)  230  62.2 3.7e-09
UNIPROT:Q6B8Y2 Q6B8Y2 Ferredoxin.                  (  98)  230  62.2 3.8e-09
UNIPROT:FER1_EQUTE P00234 Ferredoxin I.            (  95)  229  61.9 4.3e-09
UNIPROT:FER_RHOPL P07484 Ferredoxin.               (  97)  229  61.9 4.4e-09
UNIPROT:FER_GUITH O78510 Ferredoxin.               (  96)  228  61.7 5.2e-09
UNIPROT:FER_PORPU P51320 Ferredoxin.               (  98)  228  61.7 5.3e-09
UNIPROT:FER_GLEJA P00233 Ferredoxin.               (  95)  227  61.4 6.1e-09
UNIPROT:FER_MARPO P09735 Ferredoxin.               (  95)  227  61.4 6.1e-09
UNIPROT:FER1_EQUAR P00235 Ferredoxin I.            (  95)  227  61.4 6.1e-09
UNIPROT:FER_PORUM P00242 Ferredoxin.               (  98)  227  61.5 6.2e-09
UNIPROT:FER1_RAPSA P14936 Ferredoxin, root R-B1.   (  98)  226  61.2 7.4e-09
UNIPROT:O30582 O30582 Plant-type.                  (  99)  226  61.2 7.4e-09
UNIPROT:Q7XVG7 Q7XVG7 OSJNBa0073L04.7 protein.     ( 152)  227  61.5 9.1e-09
UNIPROT:FER_ODOSI P49522 Ferredoxin.               (  98)  224  60.7   1e-08
UNIPROT:FER2_RAPSA P14937 Ferredoxin, root R-B2.   (  98)  224  60.7   1e-08
UNIPROT:FER_SPIPL P00246 Ferredoxin.               (  98)  224  60.7   1e-08
UNIPROT:Q9KJL1 Q9KJL1 FdxH.                        ( 104)  224  60.7 1.1e-08
UNIPROT:Q85FT5 Q85FT5 Ferredoxin.                  (  97)  222  60.2 1.4e-08
UNIPROT:FER_BRYMA P07838 Ferredoxin.               (  98)  222  60.2 1.4e-08
UNIPROT:FER_SPIMA P00245 Ferredoxin.               (  98)  222  60.2 1.4e-08
UNIPROT:FER6_MAIZE P94044 Ferredoxin VI, chloropl  ( 155)  224  60.8 1.5e-08
UNIPROT:FER_HORVU P83522 Ferredoxin.               (  97)  221  60.0 1.7e-08

>>UNIPROT:FER_HALN1 P00216 Ferredoxin.                    (128 aa)
 initn: 870 init1: 870 opt: 870  Z-score: 1144.0  bits: 217.3 E(): 1e-55
Smith-Waterman score: 870;  100.000% identity (100.000% ungapped) in 128 aa overlap (1-128:1-128)

               10        20        30        40        50        60
HALHA  PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWP
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWP
               10        20        30        40        50        60

               70        80        90       100       110       120
HALHA  FSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHL
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO FSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHL
               70        80        90       100       110       120


HALHA  DYLQNRVI
       ::::::::
UNIPRO DYLQNRVI


>>UNIPROT:Q9YGB6 Q9YGB6 Ferredoxin.                       (129 aa)
 initn: 761 init1: 761 opt: 761  Z-score: 1001.2  bits: 190.9 E(): 9.1e-48
Smith-Waterman score: 761;  85.156% identity (85.156% ungapped) in 128 aa overlap (1-128:2-129)

                10        20        30        40        50
HALHA   PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDW
        ::::::::::.::.:::: :::.: .:.:  :::::::..:: ::::::::::::::::
```

```
UNIPRO MPTVEYLNYEVVDDNGWDMYDDDVFAEASDMDLDGEDYGSLEVNEGEYILEAAEAQGYDW
               10        20        30        40        50        60

      60        70        80        90       100       110
HALHA  PFSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKH
       ::::::::::::::.:: ::.:::::::::::::::.:.:::::::: ::::::::::::
UNIPRO PFSCRAGACANCAAIVLEGDIDMDMQQILSDEEVEDKNVRLTCIGSPDADEVKIVYNAKH
               70        80        90       100       110       120

     120
HALHA  LDYLQNRVI
       :::::::::
UNIPRO LDYLQNRVI


>>UNIPROT:FER_HALMA P00217 Ferredoxin.                    (128 aa)
 initn: 750 init1: 750 opt: 750  Z-score: 986.9  bits: 188.2 E(): 5.8e-47
Smith-Waterman score: 750;  84.375% identity (84.375% ungapped) in 128 aa overlap (1-128:1-128)

               10        20        30        40        50        60
HALHA  PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWP
       ::::::::::..::.:::: :::.: .:.:  :: ::::..:: :::::::::::::::::
UNIPRO PTVEYLNYEVVDDNGWDMYDDDVFGEASDMDLDDEDYGSLEVNEGEYILEAAEAQGYDWP
               10        20        30        40        50        60

               70        80        90       100       110       120
HALHA  FSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHL
       ::::::::::::.:: ::.:::::::::::::::.:.:::::::: ::::::::::::::
UNIPRO FSCRAGACANCAAIVLEGDIDMDMQQILSDEEVEDKNVRLTCIGSPDADEVKIVYNAKHL
               70        80        90       100       110       120


HALHA  DYLQNRVI
       ::::::::
UNIPRO DYLQNRVI


>>UNIPROT:FER_SYNP4 P15788 Ferredoxin.                    (98 aa)
 initn: 228 init1: 228 opt: 271  Z-score: 361.6  bits: 72.1 E(): 3.9e-12
Smith-Waterman score: 271;  46.739% identity (48.864% ungapped) in 92 aa overlap (32-120:7-97)

             10        20        30           40        50
HALHA  TVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYG---TMEVAEGEYILEAAEAQGYD
                                    : .:..:   :.:: . ::::..:: .: :
UNIPRO                          ASYKVTLINEEMGLNETIEVPDDEYILDVAEEEGID
                                        10        20        30

       60        70        80        90       100       110
HALHA  WPFSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAK
        :.::::::..::. .:::::::.. :..:.:...:   : :::.. ::.: . :... .
UNIPRO LPYSCRAGACSTCAGKIKEGEIDQSDQSFLDDDQIEAGYV-LTCVAYPASDCTIITHQEE
         40        50        60        70         80        90

      120
HALHA  HLDYLQNRVI
       .:
UNIPRO ELY

.....................................................................
.......................  Material Deleted  ..........................
.....................................................................


>>UNIPROT:FER_SPIMA P00245 Ferredoxin.                    (98 aa)
 initn: 191 init1: 191 opt: 222  Z-score: 297.5  bits: 60.2 E(): 1.4e-08
Smith-Waterman score: 222;  45.070% identity (45.714% ungapped) in 71 aa overlap (39-109:17-86)

       10        20        30        40        50        60
HALHA  ETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWPFSCRAGAC
                                    :..  .  :::.:::  : : :.:::::::
UNIPRO               ATYKVTLISEAEGINETIDCDDDTYILDAAEEAGLDLPYSCRAGAC
                            10        20        30        40

       70        80        90       100       110       120
HALHA  ANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHLDYLQNRVI
       ..::. .  : ::.. :..:.:...:   : :::.. :.:
UNIPRO STCAGKITSGSIDQSDQSFLDDDQIEAGYV-LTCVAYPTSDCTIQTHQEEGLY
         50        60        70         80        90

>>UNIPROT:FER6_MAIZE P94044 Ferredoxin VI, chloroplast p  (155 aa)
```

```
 initn: 212 init1: 176 opt: 224  Z-score: 297.1  bits: 60.8 E(): 1.5e-08
Smith-Waterman score: 224;  39.394% identity (41.053% ungapped) in 99 aa overlap (23-121:59-153)

                     10        20        30        40        50
HALHA          PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAA
                                        ...:.  .: :: ..  .:. .  ::::::
UNIPRO NTLSFAGHARQAARASGPRLSSRFVASAAAVLHKVKLVGPDGTEH-EFEAPDDTYILEAA
       30        40        50        60        70        80

             60        70        80        90       100       110
HALHA  EAQGYDWPFSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVK
       :. : . :::::::.:..::. .. ::.:..  ..:.: .. :  . ::::. : :: :
UNIPRO ETAGVELPFSCRAGSCSTCAGRMSAGEVDQSEGSFLDDGQMAEGYL-LTCISYPKADCV-
        90       100       110       120       130       140

            120
HALHA  IVYNAKHLDYLQNRVI
        ... :. :
UNIPRO -IHTHKEEDLY
          150

>>UNIPROT:FER_HORVU P83522 Ferredoxin.                    (97 aa)
 initn: 195 init1: 195 opt: 221  Z-score: 296.3  bits: 60.0 E(): 1.7e-08
Smith-Waterman score: 221;  47.222% identity (47.887% ungapped) in 72 aa overlap (40-111:16-86)

      10        20        30        40        50        60
HALHA  TLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWPFSCRAGACA
                                     .:: .  :::. :: .: : :.:::::.:.
UNIPRO                ATYKVKLVTPEGEVELEVPDDVYILDQAEEEGIDLPYSCRAGSCS
                             10        20        30        40

      70        80        90       100       110       120
HALHA  NCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHLDYLQNRVI
       .::. .  ::::.. :..:.:...::  : ::: . : .: :
UNIPRO SCAGKLVSGEIDQSDQSFLDDDQMEEGWV-LTCAAYPKSDVVIETHKEEELTA
          50        60        70         80        90


128 residues in 1 query   sequences
501934690 residues in 1568424 library sequences
 Tcomplib [34t23] (4 proc)
 start: Tue Sep 28 13:05:09 2004 done: Tue Sep 28 13:06:32 2004
 Total Scan time: 272.383 Total Display time:  0.033

Function used was FASTA [version 3.4t23 March 18, 2004]
```

8.1.1.3 FASTA format

The textual output as shown above is only one possible output available. In addtition to the textual output, you can request an MVIEW (a mulitple alignment view) as in Figure 8.1 or a visual fasta view (a graphical version of the signficance) as in Figure 8.2.

The textual output from the FASTA search begins with some informational messages. This includes the reference that you should cite, the version number of the program and the libraries that were searched. In this case, an optional histogram (lying on its side) has been requested of the number of sequences found with various scores. Each equal symbol in this histogram is an indicator of 2608 sequences and the asterisk indicates the expected number. The tail of the distribution is expanded in the inset. Here each equal symbol represents 13 sequences. This histogram gives you an indication of how similar the query sequence is to some of the database sequences. For a query sequence that has found a significant match, it should be well out of the tail of the distribution. In this example there are many sequences with scores larger than 120 and they are more frequent than expected by chance. These are related ferredoxin sequences from other species.

Next comes some information about the size of the database searched (note the size of the numbers) and some statistics about the search. Next comes a section that lists the sequences (along with their locus names) that have the best scores. Finally there is a section that lists the alignments that have been found by the program.

To carry out a database search in this manner, the algorithm first establishes a table containing words from the database sequences of variable length (e.g. ATCGGA, ACCCTG, GTCACA, ... for nucleotides or MK, RS, CP, ... for proteins). This type of preprocessing of the entire database is necessary to speed the subsequent search. This table is then sorted in alphabetical order and allows matching words (from the query sequences) to found rapidly. The length of these words is

Figure 8.1: The MVIEW output from http://www.ebi.ac.uk/fasta for the ferredoxin data

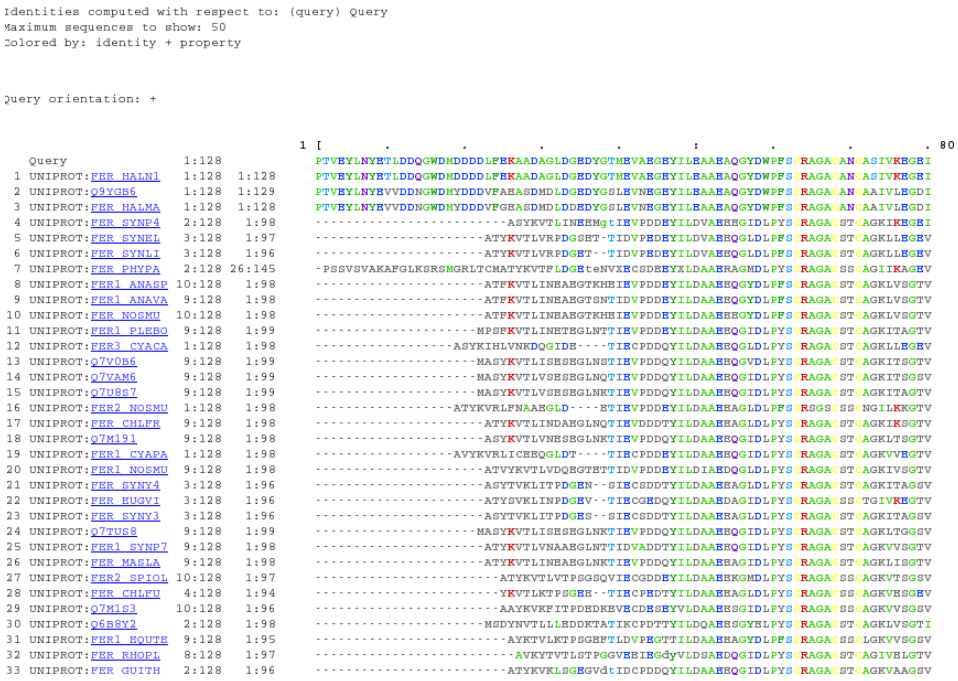
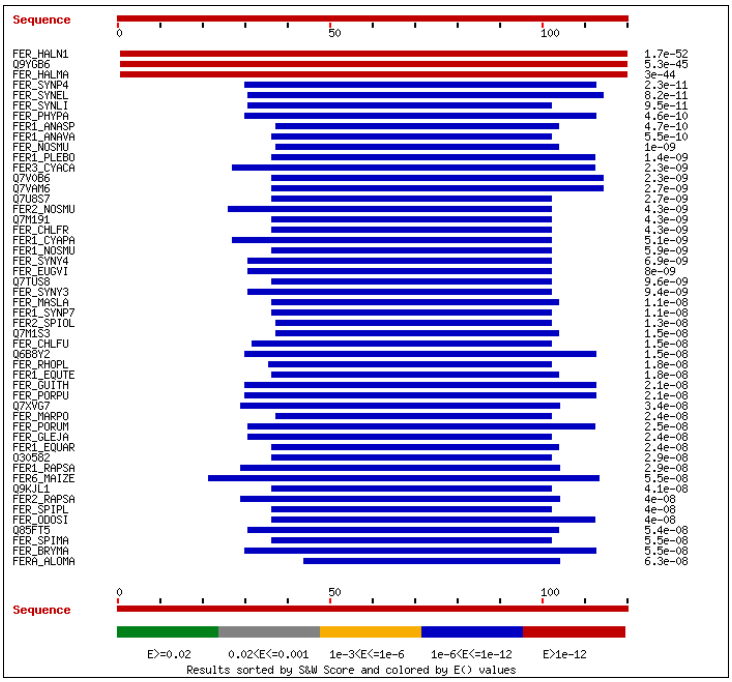


Figure 8.2: The VISUALFASTA output from http://www.ebi.ac.uk/fasta for the ferredoxin data



set by the WORD or k-tuple parameter value. By default it is 6 for nucleic acids and 2 for amino acid searches. A lower k-tuple will give a more sensitive search but will take much longer. Although a range of 3 to 6 is permitted for nucleic acids a lower value is generally unnecessary. All places in the query sequence are determined where the k-tuple from both sequences agree perfectly. Then those regions with the highest density of these identities are found.

In comparing a query sequence to the database three scores are calculated for each and every entry in the database. These scores are init1, initn and opt. An init1 score is assigned to each of these regions of high similarity after the regions are extended at the ends to include regions shorter than the length of a k-tuple and after using a BLOSUM50 matrix (alternative distance matrices are available – more on these later) to score mismatches.

Groups of larger regions are attempted to be joined together and an initn score is generated from these. This is done by setting initn equal to the sum of the two init1 scores for each region (the final init1 score of a sequence is the maximum init1 score from all interior regions). A constant of 20 is then subtracted as a joining penalty. If the initn score is less than one of the init1 scores it is discarded, the regions are not joined and the initn score will be equal to the maximum init1 score (hence initn is greater than or equal to init1).

Sequences that have an initn score larger than a cutoff value (usually 50 but this can be altered with a "LIST n" command in the query file) are then used for a Smith-Waterman alignment (see the section on alignments) and an opt score is generated from these alignments. Only the region considered significant by the program is displayed. In these alignments, the name of the sequence will be presented, the scores, and the percent similarity over the region aligned. In general the length of the region aligned is a better indicator of homology than is the percent similarity. This is because large percentages can be found in short regions just by chance. A ':' is used to indicate a complete match, a '.' to indicate a conservative amino acid replacement, and a '-' to indicate a deletion/insertion.

Note that the opt score can be lower than the initn score. This will happen when one sequence has two (or more) regions of high similarity separated by regions that have little/no homology. The two regions are joined with high init1 scores and the initn score is high because the gap penalty/join penalty is not sufficiently large. In contrast sequences with a large number of poorly similar regions will have low init1 scores but high initn scores and then low opt scores. In general, unless a very short sequence is used, the init1 score should be much improved by the opt score for truly significant sequences. Lastly a z-score based on estimates of the statistical significance of the opt scores is presented. This estimates the probability of obtaining opt scores as good or better by chance between unrelated sequences (see below).

Remember to remove repetitive sequences from your query otherwise you will get a lot of false hits. The FASTA program itself can be obtained via anonymous `ftp` if desired.

8.1.1.4 Statistical Significance

Since version 2.0 of the FASTA program distribution, FASTA, TFASTA, and SSEARCH will provide estimates of statistical significance for library searches. Work by Altschul, Arratia, Karlin, Mott, Waterman, and others (see Altschul *et al.* 1994 Nature Genetics 6:119-129 for an excellent review) shows that local sequence similarity scores follow an extreme value distribution. The probability of a database match score larger than x arising by chance alone is therefore

$$P(s \geq x) = 1 - e^{-e^{-\lambda(x-u)}}$$

where for ungapped alignments

$$u = \frac{\ln(Kmn)}{\lambda}$$

and m, n are the lengths of the query and library sequence and K and λ are constants that depend on the substitution scores and the sequence compositions. This formula can be rewritten as:

$$1 - e^{-Kmn(e^{-\lambda x})}$$

which shows that the probability of observing larger scores for unrelated library sequences increases logarithmically with the length of the library sequence (Pearson - FASTA documentation).

FASTA and SSEARCH produce gapped alignments and hence use a simple linear regression against the log of the library sequence length to calculate a normalized "z-score" with mean 50, regardless of library sequence length, and variance 10.

These z-scores can then be used with the extreme value distribution and the poisson distribution (to account for the fact that each library sequence comparison is an independent test) to calculate the expected number of library sequences required to obtain a score greater than or equal to the score obtained in the search (Pearson - FASTA documentation).

The expected number of sequences is plotted in the histogram using an '$*$'. Since the parameters for the extreme value distribution are not calculated directly from the distribution of similarity scores, the pattern of '$*$'s in the histogram gives a qualitative view of how well the statistical theory fits the similarity scores calculated by FASTA and SSEARCH. For FASTA, optimized scores are calculated for each sequence in the database and the agreement between the actual distribution of "z-scores" and the expected distribution based on the length dependence of the score and the extreme value distribution is usually very good. Likewise, the distribution of SSEARCH Smith-Waterman scores typically agrees closely with the actual distribution of "z-scores." The agreement with unoptimized scores, $ktup = 2$, is often not very good, with too many high scoring sequences and too few low scoring sequences compared with the predicted relationship between sequence length and similarity score. In those cases, the expectation values may be overestimates (Pearson - FASTA documentation).

The statistical routines assume that the library contains a large sample of unrelated sequences. If this is not the case, then the expectation values are meaningless. Likewise, if there are fewer than 20 sequences in the library, the statistical calculations are not done (Pearson - FASTA documentation).

The online FASTA - nucleotide / FASTA - protein help at EBI can be consulted for further information.

8.1.2 BLAST

While FASTA is a sensitive and rapid algorithm to search for similar sequences in the database it is not without problems. Because its initial step looks for perfect matches it might be less sensitive to more distantly related sequences that have functional homology but no longer retain complete identity. If an amino acid sequence has had many conserved replacements but no longer has identities then the FASTA algorithm might not identify these as well as it should. Fortunately, alignments where there are extensive regions of low but not exact similarity are rare enough that a small WORD or k-tuple size will pick up most regions.

A different algorithm which improves upon FASTA in speed is termed BLAST (Basic Local Alignment Search Tool). This began with a statistical paper by Karlin and Altschul (PNAS 87:2264-2268, 1990) who developed a rigorous method to obtain the probabilities of matches with a query sequence given that no gaps are permitted. This permits the use of larger WORD or k-tuple sizes with the concomitant increase in speed but permitting inexact matches between WORDs. The statistical developments permit this to be done without loss of sensitivity and allow rigorous statistical statements to be made about the matches found.

As a result of these developments Altschul, Gish, Miller, Myers and Lipman (J.Mol.Biol. 215:403-415, 1990) created the BLAST group of programs. These algorithms find ungapped, locally optimal sequence alignments. There are several versions of the BLAST programs. Some are

`BLASTN` - nucleotide query of the nucleotide database.

`BLASTP` - protein query of the protein database.

`BLASTX` - translate DNA to protein and query protein database.

`TBLASTN` - protein query of the translated nucleotide database.

`TBLASTX` - translate DNA to protein and query the translated nucleotide database.

`PHI-BLAST` - pattern-hit initiated program takes a user search pattern and finds proteins similar.

`PSI-BLAST` - use position-specific iterative score matrices to search for protein "motifs" or "profiles".

`MEGA-BLAST` - nucleotide query of the nucleotide database.

`discontiguous MEGA-BLAST` - nucleotide query of the nucleotide database.

The last two use a different algorithm than does `BLASTN`. The program `MEGA-BLAST` uses a "greedy algorithm" for nucleotide sequence alignment search and is designed to find sequences that differ slightly from the query sequence. Hence is best at identifying something "similar" in the database without concern about distant homologies. It is much faster than `BLASTN` and by default uses a much large k-tuple. The program `discontiguous MEGA-BLAST` increases

sensitivity to diverged sequences by using a discontiguous word as the initial match from which extensions are performed (see below).

To carry out this type of search go to the NCBI BLAST web server, select the desired program and fill out the forms.

Most of the options will take standard default values. The database for example, has a default of "nr". This means that it will search the non-redundant database (it includes sequences from PDB, GenBank, GenBank updates, EMBL and EMBL updates or sequences from PDB, SWISS-PROT, PIR, GenPept and GenPept updates) but there are many others that can be chosen instead. In addition you can chose to search only specific groups of organisms or to search sequences that originated from only one organism. Filter's will mask parts of your query so that things like repetitive elements are ignored (filter seq - will exclude regions of low compositional complexity, filter dust - is a modernized filter version that at the time of this writing has not yet been described in the literature. Other filter's will exclude regions with repetitive elements). It is also possible to select the number of DESCRIPTIONS n, the number of described matching sequences [100]. ALIGNMENTS n, number of high scoring pairs [50], the EXPECT n, the score such that n sequences should be found by chance alone [10] (a fractional value of one or less will give only output which is statistically unusual, larger values give more output) and the WORD size used for initial matches. Other options are available.

More information about the programs and their output can be obtained from NCBI's BLAST site including a

- BLAST overview
- BLAST FAQs
- BLAST Program Selection Guide
- BLAST course
- BLAST video tutorials
- BLAST handbook

The BLAST programs themselves can be obtained if desired by anonymous `ftp` to NCBI (with more options possible (and permissible)) and if desired, a network client that works directly through TCP/IP connections (hence, no web browser required) can be obtained as BLASTcl3 from the ftp site.

8.1.2.1 BLAST output

Typical BLAST output appears as in Figure 8.3 (this search was done on Jan 19 2002 with an APRT gene from *Mus pahari* as the query).

Each of the blue-highlighted pieces of text are links that leads directly to the entry in the database that matches the query. There is a diagram at the top of the entry that graphically demonstrates the hits and how they align with the query sequence. It is colour coded according to the statistical level of the match. In this diagram regions of low match are in gray hatch-marks. Note that even though the query sequence is in the database, there are these hatch-marks in the first matching sequence. This is because these sequence regions contain low complexity DNA (e.g. J.C. Wootton, 1994 Comput Chem 18:269-285) that would disrupt the statistical measures of similarity and hence they have been excluded by default from the match (this behaviour can be altered ... see above).

After the listing of hits comes a section that lists the match between the query sequence and the database match.

```
gi|10442645|gb|AF279458.1|AF279458  Mus musculus Ran-binding...   157  3e-35
gi|13752160|gb|AC091473.1|AC091473  Mus musculus chromosome ...   157  3e-35
gi|13194583|gb|AF316998.1|AF316998  Mus musculus D11lgp1 gen...   157  3e-35
gi|13160934|gb|AF304466.1|AF304466  Mus musculus adipocyte c...   157  3e-35
gi|12000469|gb|AC078930.13|AC078930  Mus musculus 10 BAC 280...   157  3e-35

                                Alignments

>gi|881573|gb|U28721.1|MPU28721 Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete
           cds
          Length = 2283
```

Figure 8.3: Typical results of a BLAST search

BLASTN 2.2.1 [Apr-13-2001]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
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.

RID: 1011471299-26464-27058

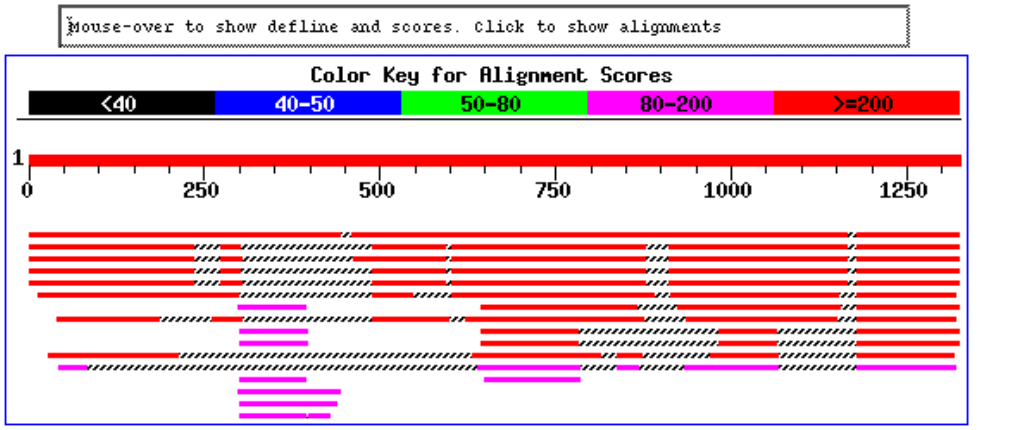
Query=
(1325 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS, or phase 0, 1 or 2 HTGS sequences)
1,074,566 sequences; 4,608,311,574 total letters

If you have any problems or questions with the results of this search
please refer to the **BLAST FAQs**

Taxonomy reports

Distribution of 2444 Blast Hits on the Query Sequence



Sequences producing significant alignments:		Score (bits)	E Value
gi|881573|gb|U28721.1|MPU28721	Mus pahari adenine phosphori...	1400	0.0
gi|881575|gb|U28720.1|MSU28720	Mus spicilegus adenine phosp...	434	e-119
gi|192011|gb|M86439.1|MUSAPRTA	Mus musculus APRT gene, part...	434	e-119
gi|192009|gb|M11310.1|MUSAPRT	Mouse adenine phosphoribosylt...	434	e-119
gi|192013|gb|M86440.1|MUSAPRTB	Mus musculus APRT gene, part...	426	e-116
gi|881577|gb|U28723.1|SLU28723	Stochomys longicaudatus aden...	333	2e-88
gi|881571|gb|U28722.1|MNU28722	Mastomys hildibrantii adenin...	276	5e-71
gi|202963|gb|L04970.1|RATAPRT	Rat adenine phosphoribosyltra...	270	3e-69
gi|13542956|gb|BC005667.1|BC005667	Mus musculus, clone MGC:...	256	4e-65
gi|12832262|dbj|AK002350.1|AK002350	Mus musculus adult male...	256	4e-65
gi|899456|gb|U28961.1|GCU28961	Gerbillus campestris adenine...	202	6e-49
gi|49607|emb|X03603.1|MAAPRTG	Hamster aprt gene for adenine...	180	2e-42
gi|17221275|emb|AL645588.7|AL645588	Mouse DNA sequence from...	176	3e-41
gi|12849531|dbj|AK012648.1|AK012648	Mus musculus 10, 11 day...	176	3e-41
gi|12845981|dbj|AK010493.1|AK010493	Mus musculus ES cells c...	176	3e-41
gi|7259320|dbj|AB032418.1|AB032418	Mus musculus mlt 1 gene,...	172	5e-40

```
 Score = 1400 bits (706), Expect = 0.0
 Identities = 706/706 (100%)
 Strand = Plus / Plus


Query: 460  gaaagaaaggtggcaagagccaccatagtggagaaggcaggtaggatccccaaggctaag 519
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1321 gaaagaaaggtggcaagagccaccatagtggagaaggcaggtaggatccccaaggctaag 1380


Query: 520  atgctaccgagtaaccatcagtgttcttctagccatagtgggcaagacctagtgttccta 579
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1381 atgctaccgagtaaccatcagtgttcttctagccatagtgggcaagacctagtgttccta 1440

.....................................................................
.......................  Material Deleted  ..........................
.....................................................................


Query: 1060 tggaggtaaagaaccagcccaagacaaacaggcttcaaagggccaggccctgtctggggt 1119
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1921 tggaggtaaagaaccagcccaagacaaacaggcttcaaagggccaggccctgtctggggt 1980


Query: 1120 gctgactaaacaaagcgcttgaataccttctctttctctgtccctt 1165
            |||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1981 gctgactaaacaaagcgcttgaataccttctctttctctgtccctt 2026

 Score =  882 bits (445), Expect = 0.0
 Identities = 445/445 (100%)
 Strand = Plus / Plus


Query: 1    aagcttgtgctaaacaactgctgtataccaggctccatgcttgagcttcagaaacaccct 60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 862  aagcttgtgctaaacaactgctgtataccaggctccatgcttgagcttcagaaacaccct 921


Query: 61   agggcagctgaatgtccaccaggagtgtccagagggagggtgagcaccccaagagaacag 120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 922  agggcagctgaatgtccaccaggagtgtccagagggagggtgagcaccccaagagaacag 981

.....................................................................
.......................  Material Deleted  ..........................
.....................................................................

Query: 361  ttcaaatcccagcaaccacatggtggctcacaaccacctacagctacagtgtacacacat 420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1222 ttcaaatcccagcaaccacatggtggctcacaaccacctacagctacagtgtacacacat 1281


Query: 421  ataataaaataaataaacaaatctt 445
            |||||||||||||||||||||||||
Sbjct: 1282 ataataaaataaataaacaaatctt 1306

 Score =  287 bits (145), Expect = 1e-74
 Identities = 145/145 (100%)
 Strand = Plus / Plus


Query: 1181 aggaaccatgtttgcagcctgtgatctgctgcaccagctacgggctgaggtggtggagtg 1240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2042 aggaaccatgtttgcagcctgtgatctgctgcaccagctacgggctgaggtggtggagtg 2101


Query: 1241 tgtgagcctggtggagctgacctcgctgaagggcagggagaggctaggacctataccatt 1300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2102 tgtgagcctggtggagctgacctcgctgaagggcagggagaggctaggacctataccatt 2161


Query: 1301 cttctctctcctccagtatgactga 1325
            |||||||||||||||||||||||||
Sbjct: 2162 cttctctctcctccagtatgactga 2186

>gi|881575|gb|U28720.1|MSU28720 Mus spicilegus adenine phosphoribosyltransferase (APRT) gene,
           complete cds
          Length = 2117

 Score =  434 bits (219), Expect = e-119
 Identities = 263/277 (94%), Gaps = 3/277 (1%)
```

```
 Strand = Plus / Plus


Query: 603  tgcctctcggctccatcccacacccttccctccttaccctaacaggtctagactccaggg 662
            |||| ||| |||||||||||||| |||||||||||||||||||||||||||||||||||
Sbjct: 1310 tgcccctcagctccatcccacaaccttccctccttaccctaacaggtctagactccaggg 1369


Query: 663  gcttcctgtttggccccttccctagctcaggagctgggcgtgggctgcgtgctcatccgga 722
            |||||||||||||||||||||||||||||||||||||||||||||||| |||||||||||||
Sbjct: 1370 gcttcctgtttggccccttccctagctcaggagctgggcgtgggctgtgctcatccgga 1429


.....................................................................
.....................  Material Deleted  ............................
.....................................................................
```

Note that this alignment might be in pieces as demonstrated above even for the database entry which is a perfect match.

Further down the listing will be generally shorter matches such as . . .

```
>gi|17221275|emb|AL645588.7|AL645588 Mouse DNA sequence from clone RP23-452K19 on chromosome 11, complete
            sequence [Mus musculus]
          Length = 5004

 Score =  176 bits (89), Expect = 3e-41
 Identities = 95/97 (97%)
 Strand = Plus / Plus


Query: 300  agagggctggtgagatggctcagcggttaggagcactgactgctcttccaaaggtcctga 359
            |||||||||||||||||||||||||||| |||||||||||||||||||||||||| ||
Sbjct: 4634 agagggctggtgagatggctcagcggttaagagcactgactgctcttccaaaggtcccga 4693


Query: 360  gttcaaatcccagcaaccacatggtggctcacaacca 396
            |||||||||||||||||||||||||||||||||||
Sbjct: 4694 gttcaaatcccagcaaccacatggtggctcacaacca 4730

>gi|12849531|dbj|AK012648.1|AK012648 Mus musculus 10, 11 days embryo whole body cDNA, RIKEN full-length
            enriched library, clone:2810002N01:related to Y39B6B.P
            PROTEIN, full insert sequence
          Length = 1026

 Score =  176 bits (89), Expect = 3e-41
 Identities = 95/97 (97%)
 Strand = Plus / Plus


Query: 302 agggctggtgagatggctcagcggttaggagcactgactgctcttccaaaggtcctgagt 361
            |||||||||||||||||||||||||| |||| ||||||||||||||||||||||||||
Sbjct: 841 agggctggtgagatggctcagcggttaagagcgctgactgctcttccaaaggtcctgagt 900


Query: 362 tcaaatcccagcaaccacatggtggctcacaaccacc 398
           ||||||||||||||||||||||||||||||||||||
Sbjct: 901 tcaaatcccagcaaccacatggtggctcacaaccacc 937

>gi|12845981|dbj|AK010493.1|AK010493 Mus musculus ES cells cDNA, RIKEN full-length enriched library,
            clone:2410015G15:related to Y39B6B.P PROTEIN, full
            insert sequence
          Length = 1022

 Score =  176 bits (89), Expect = 3e-41
 Identities = 95/97 (97%)
 Strand = Plus / Plus


Query: 302 agggctggtgagatggctcagcggttaggagcactgactgctcttccaaaggtcctgagt 361
            |||||||||||||||||||||||||| |||| ||||||||||||||||||||||||||
Sbjct: 833 agggctggtgagatggctcagcggttaagagcgctgactgctcttccaaaggtcctgagt 892
```

and finally at the bottom of the entry will be some statistics about the search . . .

```
Query: 303   gggctggtgagatggctcagcggttaggagcactgactgctctt 346
             ||||||| |||||||||||||| || | ||| |||| ||||||||
Sbjct: 30925 gggctggagagatggctcagctgtgaagaggactggctgctctt 30968

 Score = 40.1 bits (20), Expect = 4.9
 Identities = 44/52 (84%)
 Strand = Plus / Minus


Query: 345   ttccaaaggtcctgagttcaaatcccagcaaccacatggtggctcacaacca 396
             ||||| ||||||||||||  | ||||||||||||||   | ||||||||||
Sbjct: 30682 ttccagaggtcctgagttttattcccagcaaccacaccatagctcacaacca 30631

  Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,
  or phase 0, 1 or 2 HTGS sequences)
    Posted date:  Jan 19, 2002 12:06 AM
  Number of letters in database: 313,344,278
  Number of sequences in database:  1,074,566

Lambda     K      H
    1.37    0.711     1.31

Gapped
Lambda     K      H
    1.37    0.711     1.31

Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 4,147,065
Number of Sequences: 1074566
Number of extensions: 4147065
Number of successful extensions: 57560
Number of sequences better than 10.0: 2475
length of query: 1325
length of database: 4,608,311,574
effective HSP length: 22
effective length of query: 1303
effective length of database: 4,584,671,122
effective search space: 5973826471966
effective search space used: 5973826471966
T: 0
A: 30
X1: 6 (11.9 bits)
X2: 15 (29.7 bits)
S1: 12 (24.3 bits)
S2: 20 (40.1 bits)
```

8.1.2.2 BLAST format

The program output consists of three parts. The first part is a graphical diagram of the top matches to the query sequence. The second is a listing of the best matches (along with links to their database entries), their scores and their E value. The E-value is a estimate of how many matches as good or better would occur by chance alone in a database of this size. The third part is an alignment of the matches with the query sequence. The fourth part of the output will be a listing of the parameters used and some statistics of the search. Some of these parameters can be changed (see the documentation for more information) but others cannot be changed. NCBI is aware of the tradeoffs in speed versus sensitivity and attempts to offer a service with the most sensitive parameter settings that its machines can handle.

Remember that BLAST will find matches of ungapped strings. There may be more than one "ungapped" region that give an unusually large score. These multiple regions are not ignored but rather attempts are made to put them together to yield a lower overall probability. The statistics for the ungapped strings are well worked out, but the statistics for gapped matches are still not well understood.

The BLAST algorithms are capable of speeding through the entire databases within just a few seconds. Its speed is impressive. BLAST requires time proportional to the product of the query sequence length and the length of the database. The databases are growing far more quickly than are improvements in the speed of the computers or in the design of the algorithms.

The particular example shown above is a search of the database for homologues to the *Mus pahari* APRT sequence. You will note that the algorithm has done a good job at finding these homologues. The next match is the APRT gene of *Mus*

spicilegus (a closely related species – with a correspondingly closely related APRT sequence) and not surprisingly it has an expect value of 1×10^{-119}. That is, in a database of this size you expect to see 1×10^{-119} other matchs as as good or better than this one just by chance (effectively zero).

An older search of the same sequence found the same matches and if you continued down the list you would see ...

Sequences producing High-scoring Segment Pairs:		High Score	Smallest Sum Probability P(N)	N
gb|U28721|MPU28721	Mus pahari adenine phosphoribosyltr...	6451	0.0	1
gb|M86440|MUSAPRTB	Mus musculus APRT gene, partial cds.	1002	1.6e-296	10
gb|U28720|MSU28720	Mus spicilegus adenine phosphoribos...	1002	5.3e-295	12
gb|M86439|MUSAPRTA	Mus musculus APRT gene, partial cds.	1002	1.3e-290	10
gb|M11310|MUSAPRT	Mouse adenine phosphoribosyltransfe...	1002	2.6e-290	10
gb|U28723|SLU28723	Stochomys longicaudatus adenine pho...	887	5.4e-250	11
.....				
.....	Material Deleted		.....	
.....				
gb|U13835|MMCABL1	Mus musculus c-abl oncogene (c-abl)...	446	3.6e-27	2
gb|M34073|MUSMHT10C	Mus musculus (clone T10-c) MHC clas...	417	4.5e-27	2
gb|U63716|MMU63716	Mus musculus cytochrome C oxidase s...	440	4.9e-27	2
emb|Y00629|MMG37	Murine gene 37 for pot. membrane bo...	418	5.4e-27	2
dbj|D88356|D88356	Mouse DNA for 8-oxodGTPase, complet...	445	8.3e-27	2
gb|U06950|MMU06950	Mus musculus C57BL/6 lymphotoxin-be...	433	8.4e-27	2
gb|U96726|MMU96726	Mus musculus vibrator critical regi...	440	8.9e-27	2
gb|U42467|MMU42467	Mus musculus leptin receptor (Ob-r)...	432	9.6e-27	2
gb|U22062|RNU22062	Rattus norvegicus neurogranin/RC3 p...	408	1.5e-26	2
emb|X80685|MMGMCK2B	M.musculus gMCK2-beta gene	431	3.6e-25	1

So as you go down the list you find more APRT genes but also, later on, some oncogene – *c-abl*. So now you get all excited — we have discovered a new class of genes involved in cancer! Major advance ... international acclaim, ... **Nobel Prize**!! But wait, we must be cautious here, what do the statistics say. Well for this *c-abl* gene the match has a probability of 3.6×10^{-27} of occurring by chance alone. So we are home free, that is significant in anyone's statistic book. But no, life is seldom so exciting. As you continue to scan the list, you find cytochrome C, membrane proteins, growth factors, and all sorts of other genes all with apparent significant homology to the query sequence. What is going on?

Remember that BLAST (and any of the other algorithms) search for similarity not of the entire sequence but rather for any piece of the query sequence. Examining the regions of significant match between the database sequences and the query sequence indicates that these are consistently from approximately nucleotides 302 to 431 but not generally outside of this region. This region encodes a very common SINE element in rodents. Hence there is no similarity of the query gene to all these other genes but there is a significant similarity of the B2 SINE element that is inserted into the APRT gene and the B2 SINE elements that have been inserted into the other gene sequences. Be careful of the interpretation of your results — no Nobel prize this time.

Occassionally, other features such as a coiled-coil region or transmembrane regions will cause falsely positive matches to be predicted. In addition, although not a false match, the results of exon shuffling can copy a motif from one protein to another and might lead one to consider that the entire lengths of these two proteins are homologous (and derived evolutionarily from the other) when it is really only the motif that is similar. Sometimes, functional requirements will cause selection to pick on a pattern of amino acids that are similar again without homology.

Another common misuse of BLAST is to search for the most similar sequence to some query sequence. But the algorithm is designed to find similar ungapped subsequences, and to then piece these together. The order in which these sequences are ranked by score may not correspond to the order of overall similarity of the complete sequences, and certainly may not correspond to the phylogenetic history of these sequences (Koski & Golding 2001, J.Mol.Evol. 52:540-542). Thus a sequence with a higher score may not be more 'similar' to the query sequence than another sequence with a lower score (more later on what is meant by similarity). It is quite possible for the overall similarity to be greater for a sequence with a lower BLAST score. A sequence may also be more closely related in terms of history to the query than some other sequence with a lower score.

Figure 8.4: The webpage for input to an MPsrch

8.1.3 MPsrch

Discontinued

The MPsrch server at EBI runs on an HP/COMPAQ computer cluster. It uses the Smith-Waterman local similarity algorithm (see section 6.2.2 for a description of this alignment algorithm) to compare the query sequence versus the Swiss-Prot database. The advantage of this algorithm is that "is recognised as the most sensitive sequence comparison method available, whereas BLAST and FASTA utilise a heuristic one. As a consequence, MPsrch is capable of identifying hits in cases where Blast and Fasta fail and also reports fewer false-positive hits.". It will only run searches for proteins and not for nucleotides due to the time involved but also due to the discreteness of proteins. The speed achieved by MPsrch, is mainly that it is running on a "massively" parallel computer. Because of the use of a parallel computer, it was claimed that "MPsrch is the fastest implementation of the SW algorithm currently available on any machine". Many molecular biology problems lend themselves to parallel architecture computers. For many problems, intermediate steps can be effectively calculated without the need to know results from previous steps. Each of these independent steps can be given to a different processor and solved on its own. Special software has been developed for parallel computers to manage communication among individual processors and to delegate jobs to each one.

The input sequence

```
PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWP
FSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHL
DYLQNRVI
```

was given to the website of MPsrch at http://www.ebi.ac.uk/MPsrch/index.html. The input webpage for MPsrch is shown in Figure 8.4. It provides several options that you should explore. Note in particular the database search options. In the example used below I selected the database UNIPROT but for initial explorations you should try UNIREF## databases. These eliminate proteins that are within ## percentage of similarity (where ## is 100, 90 or 50). This will speed an already rapid search. .

8.1.3.1 MPsrch output

The output generated by this query is ...

```
 ***************************************************************************

        _/_/      _/_/ _/_/_/_/     _/_/_/   _/_/_/_/     _/_/_/   _/_/  _/_/
      _/_/_/  _/_/_/ _/_/  _/_/ _/_/  _/_/ _/_/  _/_/ _/_/  _/_/ _/_/  _/_/
     _/_/_/_/_/_/_/ _/_/  _/_/ _/_/       _/_/  _/_/ _/_/       _/_/  _/_/
    _/_/  _/  _/_/ _/_/_/_/     _/_/_/   _/_/_/_/   _/_/       _/_/_/_/_/
   _/_/      _/_/ _/_/             _/_/ _/_/  _/_/ _/_/       _/_/  _/_/
  _/_/      _/_/ _/_/       _/_/  _/_/ _/_/  _/_/ _/_/  _/_/ _/_/  _/_/
 _/_/      _/_/ _/_/         _/_/_/   _/_/  _/_/   _/_/_/   _/_/  _/_/  (TM)

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

                            Aneda Limited.
                      Release 4.2.80 Copyright (c)


MPsrch_pp:   protein-protein database search, using Smith-Waterman algorithm.

Run on:          Tue Aug  9 16:25:37 2005;  Search Time 4.80 secs
                                            1690.21 Million Cell updates/sec

Title:           >Sequence
Description:     No description found
Perfect Score:   1125
Sequence:        1 ptveylnyetlddqgwdmdddd.......adevkivynakhldylqnrvi  128
Scoring table:   PAM 100
                 Gap 14

Searched:        2088752 seqs, 683758041 residues

Database:        uniprot
                      1    uniprot_sprot
                      2    uniprot_trembl1
                      3    uniprot_trembl2

Statistics:      Mean 42.092;  Variance 64.372;  Scale 0.654

      Pred. No. is the number of results predicted by chance to have a
      score greater than or equal to the score of the result being printed,
      and is derived by analysis of the total score distribution.

                               SUMMARIES
                %
Result          Query
   No.  Score  Match Length DB  ID          Description            Pred. No.
 ---------------------------------------------------------------------------
     1   1125  100.0    128  1  FER_HALSA   Ferredoxin.             3.48e-265
     2    978   86.9    129  2  Q9YGB6_HALJ Ferredoxin.             4.74e-225
     3    967   86.0    128  1  FER1_HALMA  Ferredoxin 1.           4.67e-222
     4    546   48.5    138  1  FER2_HALMA  Ferredoxin 2.           4.52e-109
     5    319   28.4     98  1  FER_SYNP4   Ferredoxin.             8.96e-51
     6    305   27.1     98  1  FER1_ANAVA  Ferredoxin I.           2.59e-47
     7    302   26.8     98  1  FER1_ANASP  Ferredoxin I.           1.42e-46
     8    302   26.8     98  1  FER1_ANASO  Ferredoxin I.           1.42e-46
     9    301   26.8     97  1  FER_SYNEL   Ferredoxin I.           2.50e-46
    10    301   26.8     97  1  FER_SYNEN   Ferredoxin I.           2.50e-46
    11    301   26.8     97  1  FER_SYNVU   Ferredoxin I.           2.50e-46
    12    298   26.5     96  1  FER_SYNLI   Ferredoxin.             1.37e-45
    13    297   26.4     98  1  FER_NOSMU   Ferredoxin.             2.41e-45
    14    294   26.1     98  1  FER2_NOSMU  Ferredoxin II.          1.31e-44
    15    290   25.8     99  1  FER1_PLEBO  Ferredoxin I (FdI).     1.25e-43
    16    290   25.8     99  2  Q7DK20_PLEB Ferredoxin.             1.25e-43
    17    288   25.6     99  3  Q7U8S7_SYNP Ferredoxin.             3.84e-43
    18    286   25.4     99  3  Q7VAM6_PROM Ferredoxin.             1.18e-42
    19    284   25.2     99  3  Q7V0B6_PROM Ferredoxin.             3.64e-42
    20    282   25.1     98  2  Q7M191_SYNS Ferredoxin.             1.12e-41


                               ALIGNMENTS
   (Maximum 20 Alignments, Predicted No. Cut-off is OFF )


RESULT    1
ID   FER_HALSA      STANDARD;      PRT;   128 AA.
DE   Ferredoxin.
```

```
  DB  1;  Score    1125;  Match 100.0%;  QryMatch 100.0%;  Pred. No. 3.48e-265;
  Matches   128;  Conservative    0;  Mismatches   0;  Indels   0;  Gaps   0;

          ************************************************************
Db      1 PTVEYLNYETLDDQGWDMDDDDLFEKAADAGLDGEDYGTMEVAEGEYILEAAEAQGYDWP 60
Qy      1 ptveylnyetlddqgwdmdddlfekaadagldgedygtmevaegeyileaaeaqgydwp 60

          ************************************************************
Db     61 FSCRAGACANCASIVKEGEIDMDMQQILSDEEVEEKDVRLTCIGSPAADEVKIVYNAKHL 120
Qy     61 fscragacancasivkegeidmdmqqilsdeeveekdvrltcigspaadevkivynakhl 120

          ********
Db    121 DYLQNRVI 128
Qy    121 dylqnrvi 128


RESULT    2
ID   Q9YGB6_HALJP PRELIMINARY;      PRT;   129 AA.
DE   Ferredoxin.

  DB  2;  Score     978;  Match 85.2%;  QryMatch 86.9%;  Pred. No. 4.74e-225;
  Matches   109;  Conservative    9;  Mismatches  10;  Indels   0;  Gaps   0;

          ********* .** **** ***.*  *.*  *******..** *****************
Db      2 PTVEYLNYEVVDDNGWDMYDDDVFAEASDMDLDGEDYGSLEVNEGEYILEAAEAQGYDWP 61
Qy      1 ptveylnyetlddqgwdmdddlfekaadagldgedygtmevaegeyileaaeaqgydwp 60

          ************.** **.***************.*.********* *************
Db     62 FSCRAGACANCAAIVLEGDIDMDMQQILSDEEVEDKNVRLTCIGSPDADEVKIVYNAKHL 121
Qy     61 fscragacancasivkegeidmdmqqilsdeeveekdvrltcigspaadevkivynakhl 120

          ********
Db    122 DYLQNRVI 129
Qy    121 dylqnrvi 128


RESULT    3
ID   FER1_HALMA     STANDARD;      PRT;   128 AA.
DE   Ferredoxin 1.

  DB  1;  Score     967;  Match 84.4%;  QryMatch 86.0%;  Pred. No. 4.67e-222;
  Matches   108;  Conservative    9;  Mismatches  11;  Indels   0;  Gaps   0;

          ********* .** **** ***.*  *.*  ** ****..** *****************
Db      1 PTVEYLNYEVVDDNGWDMYDDDVFGEASDMDLDDEDYGSLEVNEGEYILEAAEAQGYDWP 60
Qy      1 ptveylnyetlddqgwdmdddlfekaadagldgedygtmevaegeyileaaeaqgydwp 60

          ************.** **.***************.*.********* *************
Db     61 FSCRAGACANCAAIVLEGDIDMDMQQILSDEEVEDKNVRLTCIGSPDADEVKIVYNAKHL 120
Qy     61 fscragacancasivkegeidmdmqqilsdeeveekdvrltcigspaadevkivynakhl 120

          ********
Db    121 DYLQNRVI 128
Qy    121 dylqnrvi 128


RESULT    4
ID   FER2_HALMA     STANDARD;      PRT;   138 AA.
DE   Ferredoxin 2.

  DB  1;  Score     546;  Match 51.7%;  QryMatch 48.5%;  Pred. No. 4.52e-109;
  Matches    61;  Conservative   27;  Mismatches  29;  Indels   1;  Gaps   1;

          **.**.* *.*.** ..*.*** ***** *.. *.*   *   . .*****  *. ***.
Db      2 VEFLNFEVLEDHGWALQDEDLFAKAADANLQSTDFGRFYVDPNDTLLEAAEKNGFAWPFA 61
Qy      3 veylnyetlddqgwdmddddlfekaadagldgedygtmevaegeyileaaeaqgydwpfs 62

          **.***.***  * .**.     .**.  *. ** .**.**..* .*. ***** ***
Db     62 CRGGACTNCAVAVVDGEMPSPASHILP-PELTEKGIRLSCIAAPVSDDAKIVYNLKHL 118
Qy     63 cragacancasivkegeidmdmqqilsdeeveekdvrltcigspaadevkivynakhl 120


RESULT    5
ID   FER_SYNP4      STANDARD;      PRT;    98 AA.
DE   Ferredoxin.

  DB  1;  Score     319;  Match 53.5%;  QryMatch 28.4%;  Pred. No. 8.96e-51;
  Matches    38;  Conservative   15;  Mismatches  17;  Indels   1;  Gaps   1;
```

```
          *.**.. ****. ** .* * *.*******. **. .******   *  * *...*   *
Db     17 TIEVPDDEYILDVAEEEGIDLPYSCRAGACSTCAGKIKEGEIDQSDQSFLDDDQIEAGYV 76
Qy     39 tmevaegeyileaaeaqgydwpfscragacancasivkegeidmdmqqilsdeeveekdv 98

           ***.. **.*
Db     77 -LTCVAYPASD 86
Qy     99 rltcigspaad 109


RESULT    6
ID   FER1_ANAVA     STANDARD;      PRT;    98 AA.
DE   Ferredoxin I.

  DB  1;  Score     305;  Match 50.7%;  QryMatch 27.1%;  Pred. No. 2.59e-47;
  Matches    38;  Conservative   16;  Mismatches  19;  Indels   2;  Gaps   2;

          *..*.. ****.*** **** *********. **. .  * .*   *  * *...*   *
Db     17 TIDVPDDEYILDAAEEQGYDLPFSCRAGACSTCAGKLVSGTVDQSDQSFLDDDQIEAGYV 76
Qy     39 tmevaegeyileaaeaqgydwpfscragacancasivkegeidmdmqqilsdeeveekdv 98

           ***.. *..* * *
Db     77 -LTCVAYPTSD-VTI 89
Qy     99 rltcigspaadevki 113


.....................................................................
.....................  Material Deleted  ............................
.....................................................................


RESULT   20
ID   Q7M191_SYNSP PRELIMINARY;      PRT;    98 AA.
DE   Ferredoxin.

  DB  2;  Score     282;  Match 47.9%;  QryMatch 25.1%;  Pred. No. 1.12e-41;
  Matches    34;  Conservative   17;  Mismatches  19;  Indels   1;  Gaps   1;

          *.**.. .***.*** ** * *.*******. **. .  * .*   *  * *...*   *
Db     17 TIEVPDDQYILDAAEEQGIDLPYSCRAGACSTCAGKLTSGTVDQSDQSFLDDDQIEAGFV 76
Qy     39 tmevaegeyileaaeaqgydwpfscragacancasivkegeidmdmqqilsdeeveekdv 98

           ***.. *..*
Db     77 -LTCVAYPTSD 86
Qy     99 rltcigspaad 109

  Search Completed:  Tue Aug  9 16:27:18 2005
Job time:  101 seconds
```

8.1.3.2 MPsrch format

This particular search took only 5 seconds of CPU time and a total of 101 seconds including input/output. This speed is a great improvement over that achieved by the FASTA algorithm. The web page output is shown in Figure 8.5. . This algorithm is as fast as BLASTP and in addition, it should also give a more sensitive search for distant homologies.

The mean and variance of the distribution of scores from the entire database are calculated. These are used to construct empirical statistics of the predicted number of random matches in the database equal to or better than that found. The algorithm then lists the best scores (50 of them here, the default for NAMES) and then lists more detailed reports for a subclass of these (30 here, the default for ALIGN). For each it calculates the raw score, the percent matches, the predicted number expected, the number of matches, the number of mismatches, the number of partial matches (residue pairs with a positive score in the PAM matrix), the number of indels and the number of gaps. This program considers these two differently in that a single gap can be composed of any number of adjacent indels.

In this case all hits have very small “pred. no.” numbers indicating that they each have statistically significant homology to the ferredoxin query sequence (not too surprising since they are all different ferredoxins). Also note that the Smith-Waterman alignment algorithm does a best local alignment (more on this later) so the entire query sequence may not be presented in the output. Rather the part of the sequence that has a good alignment with the database entry is shown. The sequence is not aligned for regions where the significance of the alignment begins to decline. Hence in the example above, for the alignment to result #20, only amino acids 17 through 86 from the database sequence and amino acids 39 to 109 from

MPsrch **Summary Table**

SUBMISSION PARAMETERS			
Title	Sequence	Database	uniprot
Sequence length	128	Sequence type	p
Program	MPsrch_pp	Version	4.2.80
Matrix	PAM 100	Open gap penalty	14
Gap extension penalty	14		

Show Annotation | MPsrch Result | XML | SUBMIT ANOTHER JOB
Show Alignments | Clear all | Check all | Invert selection | Reset

Alignment	DB:ID	Description	Length	Match%	Query Match%	Score	Pred.No.
1	UNIPROT:FER_HALSA	Ferredoxin.	128	100.0	100.0	1125	3.48e-265
2	UNIPROT:Q9YGB6_HALJP	Ferredoxin.	129	85.2	86.9	978	4.74e-225
3	UNIPROT:FER1_HALMA	Ferredoxin 1.	128	84.4	86.0	967	4.67e-222
4	UNIPROT:FER2_HALMA	Ferredoxin 2.	138	51.7	48.5	546	4.52e-109
5	UNIPROT:FER_SYNP4	Ferredoxin.	98	53.5	28.4	319	8.96e-51
6	UNIPROT:FER1_ANAVA	Ferredoxin I.	98	50.7	27.1	305	2.59e-47
7	UNIPROT:FER1_ANASP	Ferredoxin I.	98	51.4	26.8	302	1.42e-46
8	UNIPROT:FER1_ANASO	Ferredoxin I.	98	51.4	26.8	302	1.42e-46
9	UNIPROT:FER_SYNEL	Ferredoxin I.	97	50.0	26.8	301	2.50e-46
10	UNIPROT:FER_SYNEN	Ferredoxin I.	97	50.0	26.8	301	2.50e-46
11	UNIPROT:FER_SYNVU	Ferredoxin I.	97	50.0	26.8	301	2.50e-46

Figure 8.5: The webpage output from an MPsrch

the query sequence are shown in the alignment, even though the query protein is 128 amino acids in length. The sequence prior to amino acid 17/39 and after amino acid 86/109 are not considered to be part of the significant local alignment.

8.2 BLOCKS

The FASTA and BLAST servers are often searched for homologues in order to identify the query sequence. The BLOCKS server at http://blocks.fhcrc.org is designed to identify chunks of a protein that my encode some function. The BLOCKS server is thus somewhat related to the other servers mentioned above (and hence included here) but is designed to answer a different question. Instead of looking for similar sequences in the databases, it scans a database of protein motif signatures constructed from the INTERPRO database (a collection of of protein families, domains and functional sites found in known proteins that can be applied to explore unknown protein sequences). In this way, `BLOCKS` will search a query sequence (must be protein or optionally, it will translate your nucleotide sequence to a protein) for similar protein motifs in known proteins. Blocks are defined as short ungapped (but potentially with variable length) segments of highly conserved regions of proteins. As of August 2003 the `BLOCKS` database website reports that it consists of 8656 block patterns (version 13.0, Aug 2001). This search is particularly useful for analysing distantly related proteins.

The web form to search the `BLOCKS` database is located at http://blocks.fhcrc.org/blocks/blocks_search.html (References should cite S.Henikoff & J.Henikoff, 1991 Nucl.Acids.Res. 19:6565-6572). Again simply supply the web page with your query sequence.

Since this search only makes sense for proteins, if a nucleotide sequence is supplied to the server, it will be translated in all frames. But a nucleotide sequence with IUBPAC ambiguity codes will be interpreted as a protein and will remain untranslated.

8.2.1 BLOCKS output

In the example below, I have searched the BLOCKS database with the sequence

```
> Ferredoxin
GIDPNYRTHKPVVGDSSGHKIYGPVESPKVLGVHGTIVGVDFDLCIADGSCITACPVNVF
QWYETPGHPASEKKADPVNQQACIFCMACVNVCPVAAIDVKPP
```

The BLOCKS output begins with a lengthy informational message that I have deleted and then continues with the guts of the message.

```
Hits


Query=Ferredoxin n
Size=103 Amino Acids
Blocks Searched=11182
Alignments Done=         1439343
Cutoff combined expected value for hits=  1
Cutoff block expected value for repeats/other=  1
================================================================================
                                                                    Combined
Family                                          Strand  Blocks   E-value
PR00353    4Fe-4S ferredoxin signature             1   2 of 2   1.2e-06
PR00354    7Fe ferredoxin signature                1   1 of 3   0.00025
IPB000985  Legume lectin alpha domain              1   1 of 7      0.58

================================================================================
>PR00353 2/2 blocks Combined E-value= 1.2e-06: 4Fe-4S ferredoxin signature
Block    Frame    Location (aa)      Block E-value
PR00353A    0         76-87                4.5
PR00353B    0         88-99             0.00014
Other reported alignments:

                        |---  151 amino acids---|
          PR00353 AA.....................................................BB
        Ferredoxin                                              ::::::::::::AABB

PR00353A            <->A   (1,571):75
AEGA_ECOLI|P37127  80      IQVNQQKCIGCK
                             |||| || |
Ferredoxin         76      DPVNQQACIFCM

PR00353B           A<->B   (0,338):0
FER_CLOSP|P00197   42      ACANTCPVDAIV
                           || | ||| ||
Ferredoxin         88      ACVNVCPVAAID

--------------------------------------------------------------------------------
>PR00354 1/3 blocks Combined E-value= 0.00025: 7Fe ferredoxin signature
Block    Frame    Location (aa)      Block E-value
PR00354C    0         78-95             0.00027
Other reported alignments:
PR00354C    0         40-57              0.0018

                        |---   57 amino acids---|
          PR00354 AAAAA.........................BBBBB::.................CCCCCCCC
        Ferredoxin                      ::::::::::::::::::::::::::::::::CCCCCCCC
        Ferredoxin                                       CCCCCCCC

PR00354C            <->C   (34,389):77
FER_BACSC|Q45560   34      IDPDVCIDCGACEAVCPV
                                 || | ||  ||||
Ferredoxin         78      vNqqaCIfCmACVnVCPV
                   40      vDfDLCIadGsCitACPV

--------------------------------------------------------------------------------
>IPB000985 1/7 blocks Combined E-value=    0.58: Legume lectin alpha domain
Block    Frame    Location (aa)      Block E-value
IPB000985D  0         36-45                0.67
Other reported alignments:

                        |---  126 amino acids---|
         IPB000985 AAAA:...BBBBB::..CC:......DD........EEEEEE:........FFFF:..GG
        Ferredoxin                     :::::::DD

IPB000985D          <->D   (83,186):35
```

```
LECN_PEA|P16270    145     RFVGLEFDLY
                             ||  |||
Ferredoxin         36      TIVGVDFDLC

--------------------------------------------------------------------------------

3 possible hits reported
```

In this case, for ferredoxin, the program returns three possible hits. These are a 4Fe ferredoxin, a 7Fe Ferredoxin and a legume lectin alpha domain. The first signature consists of two parts (two blocks), the signature for the second hit consists of three parts (but only one was found in the query sequence) and the signature for the third hit consists of seven parts (but again only one is present in the query sequence). Each of these blocks is labelled A, B, C, etc. The E-values are calculated (as per the BLAST searches) to represent the expected number of hits with as good a similarity or better in a database of this size. Hence the last hit to a legume lection alpha domain is probably just noise.

After this initial presentation, the program returns a diagram of hits. So in the first hit, the first block (A), can typically begin anywhere from the 1^{st} to the 571^{st} amino acid (in bone-fide proteins with this signature). In our query it begins at position 75. The second block (B), can occur anywhere from 0 to 338 amino acids distant from the first block. In our query sequence it is 0 amino acids away. Alignments of each of these blocks to a best match is shown.

For the second hit, the query contains two possible locations for the "C" block but non of the other blocks. For the third hit, only the "D" block.

8.2.2 Getting the Block

In addition to this the BLOCKS server will allow you access to information about the individual blocks found. You can get the entry either via links on their web page. The following output is are examples from their links.

From the BLOCKS database itself, it has the following information on the block.

```
         Prints Database 37 in Blocks Format, Jun 2003
       Made available by the Fred Hutchinson Cancer Research Center
     1100 Fairview AV N, A1-162, PO Box 19024, Seattle, WA 98109-1024
   Based on PRINTS Database as described by TK Attwood, et al (1994),
   NAR 22(17):3590-3596. ID is from PRINTS gc line, AC is from
   PRINTS gx line, DE is from PRINTS gt line, BL is BLOCK information.
   Each PRINTS motif is represented by one block. For each segment, the
   sequence ID is followed by the position of the first residue in the
   segment. Sequence weights are shown to the right of each segment. The
   higher the weight (maximum 100) the more dissimilar the segment is from
   other segments in the block. These weights were obtained using the
   position-based method of S Henikoff & JG Henikoff (1994), JMB 243:574-578.
   Calibrated with position-specific scoring matrices made with pseudo-counts,
   JG Henikoff & S Henikoff (1996), CABIOS 12(2):135-143.
   ================================================================================


Block PR00353A

ID 4FE4SFRDOXIN; BLOCK
AC PR00353A; distance from previous block=(1,571)
DE 4Fe-4S ferredoxin signature
BL adapted; width=12; seqs=171; 99.5%=733; strength=1118
P81293             ( 275) YVIDEDLCIGCR  17
FER_CLOSP|P00197   (  30) RVIDADKCIDCG  21
O27769             (  62) VVILEDRCIGCG  41
O28894             ( 233) TYVDWDKCIGCG  30
FER_CLOAC|P00198   (  30) YVIDADTCIDCG  15
FER_BACSC|Q45560   (  32) YYIDPDVCIDCG  26
Q59575             ( 147) IEIDKDTCIYCG  18
FER2_DESDN|P00211  (   5) VIVDSDKCIGCG  21
O30081             (   6) IAIDEEKCIGCG  18
O74028             ( 147) IEIDKDTCIYCG  18
FDXH_HAEIN|P44450  ( 132) VDFQSDKCIGCG  55
O26505             ( 164) AVVDESICIGCG  26
```

```
FER_CLOTM|P07508   (  30) YVIDADACIECG  40
NUIM_CAEEL|Q22619  ( 145) YDIDMTKCIYCG  18

.....................................................................
.....................  Material Deleted .............................
.....................................................................

O29066             (   9) FVHDRRKCIGCY  81
Q03195             (  48) AFISEILCIGCG  65
FER1_RHOCA|P16021  (   2) MKIDPELCTSCG  48
O28624             (  73) LIVDESLCVGCG  20
P73811             (  77) IVIDDQSCVDCG  41
Q46606             ( 145) VVRDMGKCIRCL  78
Y719_METJA|Q58129  (  55) PVISEVLCSGCG  63
O28573             (  62) AVVNYNYCKGCG  28
O27592             ( 556) YMIDPEKCDGCM  92
P74022             ( 141) FGIDHNRCILCT  59
//

Block PR00353B

ID   4FE4SFRDOXIN; BLOCK
AC   PR00353B; distance from previous block=(0,338)
DE   4Fe-4S ferredoxin signature
BL   adapted;   width=12; seqs=171; 99.5%=728; strength=1179
P81293             ( 318) ACARECPVGAIK  11
FER_CLOSP|P00197   (  42) ACANTCPVDAIV  11
O27769             (  74) LCRDACPVGAIT  17
O28894             ( 312) PCEKACPTGAIN  13
FER_CLOAC|P00198   (  42) ACAGVCPVDAPV  15
FER_BACSC|Q45560   (  44) ACEAVCPVSAIY  17
Q59575             ( 313) ACERSCPVNAIE  11
FER2_DESDN|P00211  (  47) SCIEVCPQNAIV  20
O30081             (  18) RCVNSCPTGALV  16
O74028             ( 313) ACERSCPVTAIT  21
FDXH_HAEIN|P44450  ( 180) ACVKTCPTGAIR  12
O26505             ( 213) VCEENCPTGAIR  17
FER_CLOTM|P07508   (  42) ACANVCPVDAPQ  14

.....................................................................
.....................  Material Deleted .............................
.....................................................................

FER1_RHOCA|P16021  (  14) DCEPVCPTNAIA  29
O28624             ( 141) VCRENCPSDAIR  26
P73811             (  89) LCTGVCPTEALS  24
Q46606             ( 200) QCTLVCPVGALA  30
Y719_METJA|Q58129  (  67) ICVKRCPFKAIS  20
O28573             (  74) ICASVCPFEAIK  14
O27592             ( 568) ACIKTCPAEAIN  18
P74022             ( 197) KCVDACPTGSIF 100
//
```

This provides a short description of the parts of each block and then representative sequences that contain these blocks (with a links to that sequence, the position of the first residue in the block, the block and a weighting score). This information can be seen in graphical format as shown in Figure 8.6.

In addition you can get more data about the blocks through the PROSITE database link for this entry

```
PROSITE: PS00198
ID   4FE4S_FERREDOXIN; PATTERN.
AC   PS00198;
DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE   4Fe-4S ferredoxins, iron-sulfur binding region signature.
PA   C-x(2)-C-x(2)-C-x(3)-C-[PEG].
NR   /RELEASE=41.21,133312;
NR   /TOTAL=523(348); /POSITIVE=482(318); /UNKNOWN=2(2); /FALSE_POS=39(28);
NR   /FALSE_NEG=16; /PARTIAL=5;
CC   /TAXO-RANGE=A?EP?; /MAX-REPEAT=6;
CC   /SITE=1,iron_sulfur; /SITE=3,iron_sulfur; /SITE=5,iron_sulfur;
CC   /SITE=7,iron_sulfur;
DR   P37127, AEGA_ECOLI, T; P26474, ASRA_SALTY, T; P26476, ASRC_SALTY, T;
DR   P31894, COOF_RHORU, T; Q49161, DCA1_METMA, T; Q49163, DCA2_METMA, T;
DR   Q57617, DCMA_METJA, T; P26692, DCMA_METSO, T; O27743, DCMA_METTH, T;
DR   P08066, DHSB_BACSU, T; Q09545, DHSB_CAEEL, T; P48932, DHSB_CHOCR, T;
DR   P51053, DHSB_COXBU, T; P48933, DHSB_CYACA, T; P21914, DHSB_DROME, T;
DR   P07014, DHSB_ECOLI, T; P21912, DHSB_HUMAN, T; O42772, DHSB_MYCGR, T;
```

Figure 8.6: A map of the BLOCKS location in representative proteins

Description:	4Fe-4S ferredoxin signature
Sequences:	171
Distinct blocks:	2
Map Scaling:	|————| [100 amino acids]
Notes:	Mouse over to show start and end positions

Sequence ID	Length	Sequence
P81293	329	
FER_CLOSP|P00197	53	
O27769	85	
O28894	323	
FER_CLOAC|P00198	53	
FER_BACSC|Q45560	55	
Q59575	324	
FER2_DESDN|P00211	58	
O30081	29	
O74028	324	
FDXH_HAEIN|P44450	191	
O26505	224	
FER_CLOTM|P07508	53	
NUIM_CAEEL|Q22619	168	
FER_CLOPA|P00195	53	
FER_CLOPE|P22846	53	
Q57934	245	
Q50784	262	
O27205	262	
O27597	311	
NQO9_PARDE|P29921	119	
NUIM_ARATH|Q42599	178	
NUIM_BOVIN|P42028	168	

```
DR   Q59662, DHSB_PARDE, T; P80477, DHSB_PORPU, T; P21913, DHSB_RAT  , T;
DR   P80480, DHSB_RECAM, T; Q92JJ8, DHSB_RICCN, T; Q9ZEA1, DHSB_RICPR, T;
DR   Q8ZQU2, DHSB_SALTY, T; P21911, DHSB_SCHPO, T; P32420, DHSB_USTMA, T;

.....................................................................
.......................  Material Deleted ...........................
.....................................................................

DR   Q01642, M84A_DROME, F; Q01643, M84B_DROME, F; Q01644, M84C_DROME, F;
DR   Q01645, M84D_DROME, F; P08175, M87F_DROME, F; P55952, MT_POTPO  , F;
DR   O28002, RPOD_ARCFU, F; Q8PV16, RPOD_METMA, F; O26144, RPOD_METTH, F;
DR   Q96YW0, RPOD_SULTO, F; P23327, SRCH_HUMAN, F; P16230, SRCH_RABIT, F;
DR   P45866, YWJF_BACSU, F;
3D   1A6L; 1AXQ; 1B0T; 1B0V; 1BC6; 1BD6; 1BLU; 1BQX; 1BWE; 1C4A; 1C4C; 1CLF;
3D   1D3W; 1DUR; 1DWL; 1E7P; 1F2G; 1F5B; 1F5C; 1FCA; 1FD2; 1FDA; 1FDB; 1FDD;
3D   1FDN; 1FEH; 1FER; 1FRH; 1FRI; 1FRJ; 1FRK; 1FRL; 1FRM; 1FRX; 1FTC; 1FXD;
3D   1G3O; 1G6B; 1GAO; 1GT8; 1GTE; 1GTH; 1H7W; 1H7X; 1H98; 1HFE; 1JB0; 1K0T;
3D   1KF6; 1KFY; 1KQF; 1KQG; 1L0V; 1NEK; 1QLA; 1QLB; 1ROF; 1VJW; 1XER; 2FD2;
3D   2FDN; 5FD1; 6FD1; 6FDR; 7FD1; 7FDR;
DO   PDOC00176;
//
```

NiceSite View of PROSITE: PDOC00176 (documentation)
4Fe-4S ferredoxins, iron-sulfur binding region signature
PROSITE cross-reference(s)
PS00198; 4FE4S_FERREDOXIN
Documentation

Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron-sulfur cluster(s). One of these subgroups are the 4Fe-4S ferredoxins, which are found in bacteria and which are thus often referred as 'bacterial-type' ferredoxins. The structure of these proteins [2] consists of the duplication of a domain of twenty six amino acid residues; each of these domains contains four cysteine residues that bind to a 4Fe-4S center.

A number of proteins have been found [3] that include one or more 4Fe-4S binding domains similar to those of bacterial-type ferredoxins. These proteins are listed below (references are only provided for recently determined sequences).

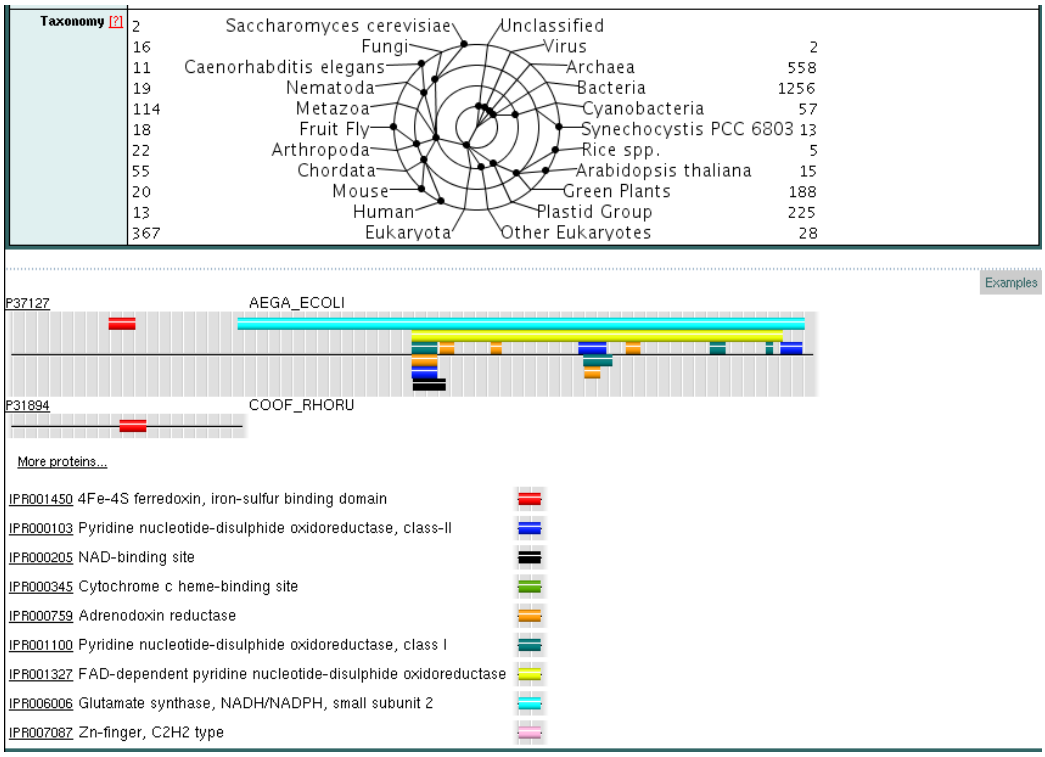
- The iron-sulfur proteins of the succinate dehydrogenase and the fumarate reductase complexes (EC 1.3.99.1). These enzyme complexes, which are components of the tricarboxylic acid cycle, each contain three subunits: a flavoprotein, an iron-sulfur protein, and a b-type cytochrome. The iron-sulfur proteins contain three different iron-sulfur centers: a 2Fe-2S, a 3Fe-3S and a 4Fe-4S.
- Escherichia coli anaerobic glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) This enzyme is composed of three subunits: A, B, and C. The C subunit seems to be an iron-sulfur protein with two ferredoxin-like domains in the N-terminal part of the protein.
- Escherichia coli anaerobic dimethyl sulfoxide reductase. The B subunit of this enzyme (gene dmsB) is an iron-sulfur protein with four 4Fe-4S ferredoxin-like domains.
- Escherichia coli formate hydrogenlyase. Two of the subunits of this oligomeric complex (genes hycB and hycF) seem to be iron-sulfur proteins that each contain two 4Fe-4S ferredoxin-like domains.
- Methanobacterium formicicum formate dehydrogenase (EC 1.2.1.2). This enzyme is used by the archaebacteria to grow on formate. The beta chain of this dimeric enzyme probably binds two 4Fe-4S centers.
- Escherichia coli formate dehydrogenases N and O (EC 1.2.1.2). The beta chain of these two enzymes (genes fdnH and fdoH) are iron-sulfur proteins with four 4Fe-4S ferredoxin-like domains.
- Desulfovibrio periplasmic [Fe] hydrogenase (EC 1.18.99.1). The large chain of this dimeric enzyme binds three 4Fe-4S centers, two of which are located in the ferredoxin-like N-terminal region of the protein.
- Methanobacterium thermoautrophicum methyl viologen-reducing hydrogenase subunit mvhB, which contains six tandemly repeated ferredoxin-like domains and which probably binds twelve 4Fe-4S centers.
- Salmonella typhimurium anaerobic sulfite reductase (EC 1.8.1.-) [4]. Two of the subunits of this enzyme (genes asrA and asrC) seem to both bind two 4Fe-4S centers.
- A Ferredoxin-like protein (gene fixX) from the nitrogen-fixation genes locus of various Rhizobium species, and one from the Nif-region of Azotobacter species.
- The 9 Kd polypeptide of chloroplast photosystem I [5] (gene psaC). This protein contains two low potential 4Fe-4S centers, referred as the A and B

```
   centers.
 - The chloroplast frxB protein which is predicted to carry two 4Fe-4S centers.
 - An ferredoxin  from a  primitive  eukaryote,  the enteric amoeba  Entamobea
   histolytica.
 - Escherichia  coli  hypothetical  protein  yjjW, a protein with a N-terminal
   region belonging to the radical activating enzymes family (see <PDOC00834>)
   and two potential 4Fe-4S centers.

The pattern of cysteine  residues in the  iron-sulfur region  is sufficient to
detect this class of 4Fe-4S binding proteins.


Description of pattern(s) and/or profile(s)
Consensus pattern
        C-x(2)-C-x(2)-C-x(3)-C-[PEG] [The four C's are 4Fe-4S ligands]
Sequences known to belong to this class detected by the pattern
        the majority of known 4Fe-4S sequences, with very few exceptions.
Other sequence(s) detected in Swiss-Prot  24.
Note in some bacterial ferredoxins, one of the two duplicated domains
        has lost one or more of the four conserved cysteines. The
        consequence of such variations is that these domains have either
        lost their iron-sulfur binding property or bind to a 3Fe-3S
        center instead of a 4Fe-4S center.
Note the last residue of this pattern in most proteins belonging
        to this group, is a Pro; the only exceptions are the Rhizobium
        ferredoxin-like proteins which have Gly, and two Desulfovibrio
        ferredoxins which have Glu. It must also be noted that the three
        non 4Fe-4S-binding proteins which are picked-up by the pattern
        have Gly in this position of the pattern.
Last update
November 1995 / Text revised.
References
[ 1]
Meyer J.
Trends Ecol. Evol. 3:222-226(1988).
[ 2]
Otaka E., Ooi T.
J. Mol. Evol. 26:257-267(1987).
[ 3]
Beinert H.
FASEB J. 4:2483-2492(1990).
[ 4]
Huang C.J., Barrett E.L.
J. Bacteriol. 173:1544-1553(1991).
[ 5]
Knaff D.B.
Trends Biochem. Sci. 13:460-461(1988).
```

This is probably more information about ferredoxin than you would ever want. But should you desire more there are links to the INTERPRO entry for this domain. In this case it is (in part)

```
InterPro 4Fe-4S ferredoxin, iron-sulfur binding domain [?] = help
IPR001450
4Fe4S_ferredoxin  Matches: 2183 proteins
View matches: [Overview][...sorted by Name][of known structure][Detailed view][Table view]
Name [?] 4Fe-4S ferredoxin, iron-sulfur binding domain
Signatures [?] PF00037;fer4 (1903 proteins)
PR00353;4FE4SFRDOXIN (3 proteins)
PS00198;4FE4S_FERREDOXIN (2111 proteins)
Type [?] Domain
Dates [?] 1999-10-08 17:07:25.0 (created)
2000-06-29 10:12:25.0 (modified)
Found in [?] IPR000813; 7Fe ferredoxin
IPR001080; 3Fe-4S ferredoxin
IPR004452; Iron-sulfur cluster binding protein
IPR004453; 4Fe-4S cluster binding
IPR004460; CO dehydrogenase/acetyl-CoA synthase complex alpha subunit
IPR004489; Succinate dehydrogenase/fumarate reductase iron-sulfur protein
IPR004494; MauM/NapG ferredoxin-type protein
IPR004496; Ferredoxin-type protein NapF
IPR004497; NADH-plastoquinone oxidoreductase, subunit I
IPR006470; Formate dehydrogenase, beta subunit
IPR006547; Nitrate reductase, beta subunit
Process [?] electron transport (GO:0006118)
Function [?] electron transporter activity (GO:0005489)
Abstract [?]

Ferredoxins are iron-sulphur proteins that mediate electron transfer
```

Figure 8.7: Typical results from an INTERPRO query




```
in a range of metabolic reactions; they fall into several subgroups
according to the nature of their iron-sulphur cluster(s) [1, 2]. One
group, originally found in bacteria, has been termed "bacterial-type",
in which the active centre is a 4Fe-4S cluster. 4Fe-4S ferredoxins
may in turn be subdivided into further groups, based on their sequence
properties. Most contain at least one conserved domain, including four
Cys residues that bind to a 4Fe-4S centre.

During the evolution of bacterial-type ferredoxins, intrasequence gene
duplication, transposition and fusion events occured, resulting in the
appearance of proteins with multiple iron-sulphur centres: e.g. dicluster-
type (2[4Fe-4S]) and polyferredoxins, iron-sulphur subunits of bacterial
succinate dehydrogenase/fumarate reductase, formate hydrogenlyase and
formate dehydrogenase complexes, pyruvate-flavodoxin oxidoreductase,
NADH:ubiquinone reductase and others. In some bacterial ferredoxins,
one of the duplicated domains has lost one or more of the four conserved
Cys residues. These domains have either lost their iron-sulphur binding
property, or bind to a 3Fe-4S centre instead of a 4Fe-4S centre. 3D
structures are now known both for a number of monocluster-type [3]
and dicluster-type [4] 4Fe-4S ferredoxins.

CAUTION: PRINTS signature in the current entry is known to miss protein
matches and should be updated in the near future.
```

There is even a link to give a graphical interpretation of the block's taxonomic diversity and graphical demonstrations of the block's location within proteins as shown in Figure 8.7.

A really great resource.

8.3 SSearch

At the extreme slow end of database searchers is SSEARCH. This does a universal sequence comparison using the Smith-Waterman algorithm (T.F. Smith and M.S. Waterman, J.Mol.Biol. 147:195-197, 1981). That is, it is completely rigorous comparison of each sequence with the query sequence. This program uses code developed by X. Huang, R.C. Hardison, W. Miller (1990 CABIOS 6:373-381) for calculating the local similarity score and code from the ALIGN program (see below) for calculating the local alignment. SSEARCH is about 100-times slower than FASTA with ktup=2 (for proteins). The program itself is available for download as part of the FASTA package of programs.

A study by Pearson (1995 Protein Science 4:1145-1160) compared the different methods of searching the protein databases. He found that the complete Smith-Waterman algorithm performed best to find distantly related homologies, followed by FASTA and then `blastp` when using suitable scoring matrices (BLOSUM55 – more on these later) and optimal gap penalties.

8.4 Why you should routinely check your sequence

The following is an example of why you should routinely do a search (FASTA, BLAST or whatever) for any new sequence that you are working on. This is a copy of a letter to the editor of NATURE.

Fact and fiction in alignment.
NATURE 358:271, 1992

Sir - We have discovered a startling similarity between a dinosaur DNA sequence reported in the novel Jurassic Park[1] *and a partial human brain cDNA sequence from the Venter laboratory described in Nature*[2] *(see figure).*

The dinosaur sequence (DINO1) consists of duplication, with 117 base pairs from the first member of the repeat aligning with the human sequence, HUMXT01431, at the 95 per cent level of identity with only two gaps. The extraordinary degree of nucleotide sequence conservation between organisms as distantly related as dinosaur and human suggests strongly conserved function. Expression of HUMXT01431 in human brain raises the possibility that the dinosaurs were smarter than has been supposed, arguing against the hypothesis that their extinction resulted from lack of intelligence.

Our discovery also seems to raise the interesting legal question as to whether the copyright on Jurassic Park takes precedence over the pending patent on the human sequence. However, it appears that neither group is entitled to legal protection for its sequence, because both sequences also align with cloning vector pBR322, raising the possibility that both groups inadvertently sequenced vector DNA.

Alan C. Christensen, *Dept of Biochemistry and Molecular Biology, Thomas Jefferson University, Philadelphia, Pennsylvania, 19107 USA.*

Steven Henikoff, *Howard Hughes Medical Institute and Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle Washington 98104 USA.*

1 Crichton, M. Jurassic Park, 102 (Ballantine, New York 1990).

2 Adams, M.D. et al., Nature 355, 632-634 (1992).

```
HUMXT 317 GCGTTGCTGGCGTTTTTCCATAGGCTCCGACCCCCTGACGAGCATCACAAAAATCGACGCTCAA
          ***************************** ******************************
DINO1   1 GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGC----
          ************************************************** *
DINO1 670 GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAACAAGTCAGA----


HUMXT 234 GTCANAGGTGGCGGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTTGGAGCTTCC
                ******* ****************************************** * **** **
DINO1  61 ------GGTGGCG-AAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCC
                ******* ******************************************** * *
DINO1 730 ------GGTGGCG-AAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCGCTC
```

With such good jokers in the world as these gentlemen are, you don't want to get caught by them.