XNU Filter

Robert Bakarić
rbakaric@irb.hr
bakaric@evolbio.mpg.de
13.10.2015
XnuFilt-0.01

Abstract

This is a C++ implementation of XNU filtering strategy proposed by Jean Michel Claverie and David J. States (1993), created for identifying and masking repetitive segments in amino acid sequences.

Contents

1 Installation

The simplest way to compile this program is to:

1. Unpack the xnufilt package (`xnufilt-XXX.tar.gz`):

   ```
   tar -xvzf xnufilt-XXX.tar.gz
   ```

2. Change the current directory to `xnufilt-XXX`:

   ```
   cd xnufilt-XXX/
   ```

3. Configure the program for your system (–bindir is optional):

   ```
   ./configure --bindir=/absolute/directory/path/xnufilt-xxx/bin
   ```

4. Compile the program:

   ```
   make
   ```

5. Install the program:

   ```
   make install
   ```

Your binaries should be located in your local bin directory if `--bindir` option has been set. Otherwise installation needs to be carried out with root privileges in order to be installed into `/usr/local/bin` directory.

2 Input files

The xnufilt takes a regular (mulit-)fasta file as input. The example can be found in `./xnufilt-xxx/demo` and it should look like this:

`hox.fa`:

```
>gi|500757|gb|AAA86954.1| HOX A1 homeodomain protein [Homo sapiens]
MDNARMNSFLEYPILSSGDSGTCSARAYPSDHRITTFQSCAVSANSCGGDDRFLVGRGVQIGSPHHHHHHHHHHPQPATY
QTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEADVSGGYPQCAPAVYSGNLSSPMVQHHHHHQGYAGGAVGSPQYI
HHSYGQEHQSLALATYNNSLSPLHASHQEACRSPASETSSPAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFTTKQ
LTELEKEFHFNKYLTRARRVEIAASLQLNETQVKIWFQNRRMKQKKREKEGLLPISPATPPGNDEKAEESSEKSSSSPCV
PSPGSSTSDTLTTSH
>Rand_Seq_Prt
NWELYLYDPAGHRIRSWVSPNPVHFYADHCPYYPIFPRNVTTQWSPDTAGWDFEAPHTKHCTTVMMRRCALPDVIRSCSG
SSFRYRKYITKAHWICVMIWNHLSANKMKMGDQPWKECHYFKHVSCMANFAHPPVGGHKECVQCMFAWGCKNWFFNHVMP
ALKCWMKPGSEFCHHHHHHHHHHHHHHHHHHHHHHHHHHHHERVCYMHNIHRNWYHGDQSYGDECILKPGIILEYVYRCVD
DCFHWWFCAKDEPHKLMSTSFPRIMCTPLMPGCIEARIMPLCWYADWLHRRQYDCWSLCKFCANNVTPHMYKLYNPQYLW
YIASNINVTDHRKLICKRWVDDPERNFGKLVSYWSGSDLSVVPRSQYPDWRNHHMNPYTNCANFYWILNYVDCNVLHRMI
YPSDMFMSAKIETRQQDDLIENLLAWWYKQKTAWMTGPRTPHFRNSTWWFHVWIYAPTRNDPANLMVCNWYGQYVYDILW
RNEGLNTVAILEKTDQMCGWAHCFPQHMCSKQSEQHNHIIIIIIIIIIIIIIIIIHHHHHDHRKTASYENKSFVISPCQ
KNGRHRKQPTQFGHCVNSMEHSGYGLVTKFVINCHRNSMWNTKWTFIWADRAPRSWSKILGVFLNYATDDERKGSDGRLW
WKELVTFLRHKAQSCWHPVWECTADQCGRTNWQGQYLMNVGVCVHHFVSDVCMLQYPFVVNGTCAVMSKWKRHRYWVCFH
MDPYMEEYHKYWSRFPEWKAFPCNRPYKSRICRMMQMWTLAVQCQVITRFHGHWAESPKTILTFHCQFGKEHVQACVDKY
HFGAKLRPTLELQLWMKVAEAKISYFKRGMAATQHDNYYCEMNSPWLSLWHFIVFVHCINWEK
>Test|test|
RWVDDPERNFGKLVSYWSGSDLSVVPRSQYPDWRNHHMNPYTNCANFYWILNYVDCNVLHRMIFHCQFGKEHVQACVDKY
YPSDMFMSAKIMSAKIMSAKIMSAKIMSAKIMSAKIGPRTPHFRNSTWWFHVWIYAPTRNDPANLMVCNWYGQYVYDILW
LEILEILEILEILEILEILEILEFPQHMCSKQSEQHNHIFRGRHFGHKTFVKPQTDDCETTDHRKTASYENKSFVISPCQ
KNGRHRKQPTQFGHCVNSMEHSGYGLVTKFVINCHRNSMWNTKWTFIWADRAPRSWSKILGVFLNYATDDERKGSDGRLW
WKELVTFLRHKAQSCWHPVWECTADQCGRTNWQGQYLMNVGVCVHHFVSDVCMLQYPFVVNGTCAVMSKWK
```

3 Program options

I order to see program options type:

```
./bin/xnufilt -h
```

Expected output:

```
Usage: ./program [options]



 __   __            ______ _ _ _
 \ \ / /           |  ____(_) | |
  \ V / _ __  _   _| |__   _| | |_
   > < | '_ \| | | |  __| | | | __|
  / . \| | | | |_| | |    | | | |_
 /_/ \_\_| |_|\__,_|_|    |_|_|\__|


                            by Robert Bakaric

_____________________________________________________________________v0.01
**********************************************************************

CONTACT:
   This program has been written and is maintained by Robert Bakaric,
   email: rbakaric@irb.hr , bakaric@evolbio.mpg.de



LICENSE:
   The program is distributed under the GNU General Public License.
   You should have received a copy of the licence together  with this
   software. If not, see http://www.gnu.org/licenses/
----------------------------------------------------------------------
**********************************************************************

Options:
  -h [ --help ]                  produce help message
  -v [ --version ]               print version information
  -i [ --input-file ] arg        input file
  -o [ --output-file ] arg       output file
  -P [ --pam ] arg               PAM  matrix to use: 60/120/250.
  -S [ --score ] arg             Score cutoff.
  -p [ --probability ] arg       Probability cutoff.
  -m [ --min_search_offset ] arg Minimum search offset.
  -M [ --max_search_offset ] arg Maximum search offset.
```

It should be noted that default values are set unless explicitly specified.

4 Example

4.1 XnuFilt.cpp

A minimal example demonstrating the usage of xnufilt demo program:

```
./bin/XnuFilt -i ./demo/hox.fa

>gi|500757|gb|AAA86954.1| HOX A1 homeodomain protein [Homo sapiens]
MDNARMNSFLEYPILSSGDSGTCSARAYPSDHRITTFQSCAVSANSCGGDDRFLVGRGVQIGSPXXXXXXXXXXXXXXAT
YQTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEADVSGGYPQCAPAVYSGNLSSPMVXXXXXXXGYAGGAVGSPQ
YIHHSYGQEHQSLALATYNNSLSPLHASHQEACRSPASETSSPAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFT
TKQLTELEKEFHFNKYLTRARRVEIAASLQLNETQVKIWFQNRRMKQKKREKEGLLPISPATPPGNXXXXXXXXXXXXS
SPCVPSPGSSTSDTLTTSH
>Rand_Seq_Prt
NWELYLYDPAGHRIRSWVSPNPVHFYADHCPYYPIFPRNVTTQWSPDTAGWDFEAPHTKHCXXXXXXXCALPDVIRSCS
GSSFRYRKYITKAHWICVMIWNHLSANKMKMGDQPWKECHYFKHVSCMANFAHPPVGGHKECVQCMFAWGCKNWFFNHV
MPALKCWMKPGSEFCXXXXXXXXXXXXXXXXXXXXXXXXXXXXXERVCYMHNIHRNWYHGDQSYGDECILKPGIILEYVYR
```

```
CVDDCFHWWFCAKDEPHKLMSTSFPRIMCTPLMPGCIEARIMPLCWYADWLHRRQYDCWSLCKFCANNVTPHMYKLYNP
QYLWYIASNINVTDHRKLICKRWVDDPERNFGKLVSYWSGSDLSVVPRSQYPDWRNHHMNPYTNCANFYWILNYVDCNV
LHRMIYPSDMFMSAKIETRQQDDLIENLLAWWYKQKTAWMTGPRTPHFRNSTWWFHVWIYAPTRNDPANLMVCNWYGQY
VYDILWRNEGLNTVAILEKTDQMCGWAHCFPQHMCSKQSXXXXXXXXXXXXXXXXXXXXXXXXXXXXXRKTASYENKS
FVISPCQKNGRHRKQPTQFGHCVNSMEHSGYGLVTKFVINCHXXXXXXXXXXXXXXDRAPRSWSKILGVFLNYATDDER
KGSDGRLWWKELVTFLRHKAQSCWHPVWECTADQCGRTNWQGQYLMNVGVCVHHFVSDVCMLQYPFVVNGTCAVMSKWK
RHRYWVCFHMDPYMEEYHKYWSRFPEWKAFPCNRPYKSRICRMMQMWTLAVQCQVITRFHGHWAESPKTILTFHCQFGK
EHVQACVDKYHFGAKLRPTLELQLWMKVAEAKISYFKRGMAATQHDNYYCEMNSPWLSLWHFIVFVHCINWEK
>Test|test|
RWVDDPERNFGKLVSYWSGSDLSVVPRSQYPDWRNHHMNPYTNCANFYWILNYVDCNVLHRMIFHCQFGKEHVQACVDK
YYPSDMFMSAKIMSAKIMSAKIMSAKIMSAKIMSAKIGPRTPHFRNSTWWFHVWIYAPTRNDPANLMVCNWYGQYVYDI
LWXXXXXXXXXXXXXXXXXXXXXXFPQHMCSKQSEQHNHIFRGRHFGHKTFVKPQTDDCETTDHRKTASYENKSFVIS
PCQKNGRHRKQPTQFGHCVNSMEHSGYGLVTKFVINCHXXXXXXXXXXXXXXDRAPRSWSKILGVFLNYATDDERKGSD
GRLWWKELVTFLRHKAQSCWHPVWECTADQCGRTNWQGQYLMNVGVCVHHFVSDVCMLQYPFVVNGTCAVMSKWK
```

4.2 XNU.hpp

Adding the `xnufilt.hpp` header file to your program will allow you to include all the functions described in section 4. A minimal example:

```
    #include<string>
    #include<XNU.hpp>

    string ProtSeq = "\
VGRGVQIGSPHHHHHHHHHPQPATYQTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYAL\
NQEADVSGGYPQCAPAVYSGNLSSPMVQHHHHHQGYAGGAVGSPQYIHHSYGQEHQSLA\
LATYN";

/* Make object */

    /* Construction */
    XNU<int> Xnu;
    /*  OR  */
    XNU<int> Xnu(arg); // arg is : unordered_map<string, string>


/*  Functions */


    string mask = Xnu.Filter(ProtSeq);
    /* mask = VGRGVQIGSPXXXXXXXXXXXXXATYQTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYAL
     *        NQEADVSGGYPQCAPAVYSGNLSSPMVXXXXXXXGYAGGAVGSPQYIHHSYGQEHQSLA
     *        LATYN
     */
```

5 Acknowledgement

Jean Michel Claverie & David J. States (1993) Computers and Chemistry 17: 191-201.

6 Future work

Upon request!