FastaPlus

Robert Bakarić
rbakaric@irb.hr
bakaric@evolbio.mpg.de
10.11.2015
FastaPlus-0.01

Abstract

This library is designed to provide a set of easy-to-use data formatting utilities for (multi)fasta formatted files. It facilitates (DNA:to be implemented)AA sequence cleaning strategies like low complexity segment filtering (both SEG[1] and XNU[2]) and dubious character replacement. Moreover, the library enables a unique indexing of individual records based on their taxonomy identifier as well as some particular substring position.

Contents

1 Installation

The simplest way to compile this program is to:

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

   ```
   tar -xvzf fastaplus-XXX.tar.gz
   ```

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

   ```
   cd fastaplus-XXX/
   ```

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

   ```
   ./configure --bindir=/absolute/directory/path/fastaplus-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 TestFastaPlus

The program can serves as a "sandbox" utility that can be changes accordingly.

2.1 Input files

The requires a simple (multi) fasta formated file an example of which can be found in `./fastaplus-xxx/demo` directory and should look like this:

`Saccharomyces_cerevisiae.fa`:

```
>pgi|109800001|ti|559292|pi|0| YAL069W pep:known chromosome:SacCer_Apr2011:I:335:649:1
gene:YAL069W transcript:YAL069W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MIVNNTHVLTLPLYTTTTCHTHPHLYTDFTYAHGCYSIYHLKLTLLSDSTSLHGPSLTESVPNALTSLCTALASAVYTL
CHLPITPIIIHILISISHSAVPNIV
>pgi|109800002|ti|559292|pi|0| YAL068W-A pep:known chromosome:SacCer_Apr2011:I:538:792:1
gene:YAL068W-A transcript:YAL068W-A description:"Dubious open reading frame unlikely to
encode a protein\x3b identified by gene-trapping, microarray-based expression analysis, and
genome-wide homology searching"
MHGTCLSGLYPVPFTHNAHHYPHFDIYISFGGPKYCITALNTYVIPLLHHILTTPFIYTYVNITEKSPQKSPKHKNILL
FNNNT
>pgi|109800003|ti|559292|pi|0| YAL068C pep:known chromosome:SacCer_Apr2011:I:1807:2169:-1
gene:YAL068C transcript:YAL068C description:"Protein of unknown function, member of the
seripauperin multigene family encoded mainly in subtelomeric regions"
MVKLTSIAAGVAAIAATASATTTLAQSDERVNLVELGVYVSDIRAHLAQYYMFQAAHPTETYPVEVAEAVFNYGDFTTM
LTGIAPDQVTRMITGVPWYSSRLKPAISSALSKDGIYTIAN
>pgi|109800004|ti|559292|pi|0| YAL067W-A pep:known chromosome:SacCer_Apr2011:I:2480:2707:1
gene:YAL067W-A transcript:YAL067W-A description:"Putative protein of unknown function\x3b
identified by gene-trapping, microarray-based expression analysis, and genome-wide
homology searching"
MPIIGVPRCLIKPFSVPVTFPFSVKKNIRILDLDPRTEAYCLSLNSVCFKRLPRRKYFHLLNSYNIKRVLGVVYC
```

```
>pgi|109800005|ti|559292|pi|0| YAL067C pep:known chromosome:SacCer_Apr2011:I:7235:9016:-1
gene:YAL067C transcript:YAL067C description:"Putative permease, member of the allantoate
transporter subfamily of the major facilitator superfamily\x3b mutation confers resistance
to ethionine sulfoxide"
MYSIVKEIIVDPYKRLKWGFIPVKRQVEDLPDDLNSTEIVTISNSIQSHETAENFITTTSEKDQLHFETSSYSEHKDNV
NVTRSYEYRDEADRPWWRFFDEQEYRINEKERSHNKWYSWFKQGTSFKEKKLLIKLDVLLAFYSCIAYWVKYLDTVNIN
NAYVSGMKEDLGFQGNDLVHTQVMYTVGNIIFQLPFLIYLNKLPLNYVLPSLDLCWSLLTVGAAYVNSVPHLKAIRFFI
GAFEAPSYLAYQYLFGSFYKHDEMVRRSAFYYLGQYIGILSAGGIQSAVYSSLNGVNGLEGWRWNFIIDAIVSVVVGLI
GFYSLPGDPYNCYSIFLTDDEIRLARKRLKENQTGKSDFETKVFDIKLWKTIFSDWKIYILTLWNIFCWNDSNVSSGAY
LLWLKSLKRYSIPKLNQLSMITPGLGMVYLMLTGIIADKLHSRWFAIIFTQVFNIIGNSILAAWDVAEGAKWFAFMLQC
FGWAMAPVLYSWQNDICRRDAQTRAITLVTMNIMAQSSTAWISVLVWKTEEAPRYLKGFTFTACSAFCLSIWTFVVLYF
YKRDERNNAKKNGIVLYNSKHGVEKPTSKDVETLSVSDEK
>pgi|109800006|ti|559292|pi|0| YAL066W pep:known chromosome:SacCer_Apr2011:I:10091:10399:1
gene:YAL066W transcript:YAL066W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MLSLVKRSILHSIPITRHILPIQLILVKMNHVQIRNIKLYHFISYGFMLTKLTVFLFNLFFYRLRILCRLTLLILSLPV
QIYIKEIQTKMLEKHTASDTSCI
>pgi|109800007|ti|559292|pi|0| YAL065C pep:known chromosome:SacCer_Apr2011:I:11565:11951:-1
gene:YAL065C transcript:YAL065C description:"Putative protein of unknown function\x3b has
homology to FLO1\x3b possible pseudogene"
MNSATSETTTNTGAAETTTSTGAAETKTVVTSSISRFNHAETQTASATDVIGHSSSVVSVSETGNTKSLITSGLSTMSQ
QPRSTPASSIIGSSTASLEISTYVGIANGLLTNNGISVFISTVLLAIVW
```

2.2 Program options

I order to see program options type:

```
./bin/FastaPlusTest -h
```

Expected output:

```
Usage: ./program [options]



  ______        _        _____  _
 |  ____|      | |      |  __ \| |
 | |__ __ _ ___| |_ __ _| |__) | |_   _ ___
 |  __/ _` / __| __/ _` |  ___/| | | | / __|
 | | | (_| \__ \ || (_| | |    | | |_| \__ \
 |_|  \__,_|___/\__\__,_|_|    |_|\__,_|___/

                            by Robert Bakaric

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

CONTACT:
   Program Fasta 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
  -t [ --taxid ] arg       taxonomy identifier
  -o [ --output-file ] arg output file
```

2.3 Example

A minimal example demonstrating the usage of FastaPlusTest program:

```
./bin/FastaPlusTest -i demo/Saccharomyces\_cerevisiae.fa -t 12345

This file contains:6692 sequences
This file contains:3010216 sequence characters
I added this sequence to my container: HI MY NAME IS ROBERT and this Is my Se
My name is: ROBERT
And you can locate my sequence in out directory under 00000000000000000001000
```

3 SplitFasta

The program demonstrates a simple multi fasta file split procedure based on FastaPlus library.

3.1 Input files

The requires a simple (multi) fasta formated file an example of which can be found in `./fastaplus-xxx/demo` directory and should look like this:

`Saccharomyces_cerevisiae.fa`:

```
>pgi|109800001|ti|559292|pi|0| YAL069W pep:known chromosome:SacCer_Apr2011:I:335:649:1
gene:YAL069W transcript:YAL069W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MIVNNTHVLTLPLYTTTTCHTHPHLYTDFTYAHGCYSIYHLKLTLLSDSTSLHGPSLTESVPNALTSLCTALASAVYTL
CHLPITPIIIHILISISHSAVPNIV
>pgi|109800002|ti|559292|pi|0| YAL068W-A pep:known chromosome:SacCer_Apr2011:I:538:792:1
gene:YAL068W-A transcript:YAL068W-A description:"Dubious open reading frame unlikely to
encode a protein\x3b identified by gene-trapping, microarray-based expression analysis, and
genome-wide homology searching"
MHGTCLSGLYPVPFTHNAHHYPHFDIYISFGGPKYCITALNTYVIPLLHHILTTPFIYTYVNITEKSPQKSPKHKNILL
FNNNT
>pgi|109800003|ti|559292|pi|0| YAL068C pep:known chromosome:SacCer_Apr2011:I:1807:2169:-1
gene:YAL068C transcript:YAL068C description:"Protein of unknown function, member of the
seripauperin multigene family encoded mainly in subtelomeric regions"
MVKLTSIAAGVAAIAATASATTTLAQSDERVNLVELGVYVSDIRAHLAQYYMFQAAHPTETYPVEVAEAVFNYGDFTTM
LTGIAPDQVTRMITGVPWYSSRLKPAISSALSKDGIYTIAN
>pgi|109800004|ti|559292|pi|0| YAL067W-A pep:known chromosome:SacCer_Apr2011:I:2480:2707:1
gene:YAL067W-A transcript:YAL067W-A description:"Putative protein of unknown function\x3b
identified by gene-trapping, microarray-based expression analysis, and genome-wide
homology searching"
MPIIGVPRCLIKPFSVPVTFPFSVKKNIRILDLDPRTEAYCLSLNSVCFKRLPRRKYFHLLNSYNIKRVLGVVYC
>pgi|109800005|ti|559292|pi|0| YAL067C pep:known chromosome:SacCer_Apr2011:I:7235:9016:-1
gene:YAL067C transcript:YAL067C description:"Putative permease, member of the allantoate
transporter subfamily of the major facilitator superfamily\x3b mutation confers resistance
to ethionine sulfoxide"
MYSIVKEIIVDPYKRLKWGFIPVKRQVEDLPDDLNSTEIVTISNSIQSHETAENFITTTSEKDQLHFETSSYSEHKDNV
NVTRSYEYRDEADRPWWRFFDEQEYRINEKERSHNKWYSWFKQGTSFKEKKLLIKLDVLLAFYSCIAYWVKYLDTVNIN
NAYVSGMKEDLGFQGNDLVHTQVMYTVGNIIFQLPFLIYLNKLPLNYVLPSLDLCWSLLTVGAAYVNSVPHLKAIRFFI
GAFEAPSYLAYQYLFGSFYKHDEMVRRSAFYYLGQYIGILSAGGIQSAVYSSLNGVNGLEGWRWNFIIDAIVSVVVGLI
GFYSLPGDPYNCYSIFLTDDEIRLARKRLKENQTGKSDFETKVFDIKLWKTIFSDWKIYILTLWNIFCWNDSNVSSGAY
LLWLKSLKRYSIPKLNQLSMITPGLGMVYLMLTGIIADKLHSRWFAIIFTQVFNIIGNSILAAWDVAEGAKWFAFMLQC
FGWAMAPVLYSWQNDICRRDAQTRAITLVTMNIMAQSSTAWISVLVWKTEEAPRYLKGFTFTACSAFCLSIWTFVVLYF
YKRDERNNAKKNGIVLYNSKHGVEKPTSKDVETLSVSDEK
>pgi|109800006|ti|559292|pi|0| YAL066W pep:known chromosome:SacCer_Apr2011:I:10091:10399:1
gene:YAL066W transcript:YAL066W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MLSLVKRSILHSIPITRHILPIQLILVKMNHVQIRNIKLYHFISYGFMLTKLTVFLFNLFFYRLRILCRLTLLILSLPV
QIYIKEIQTKMLEKHTASDTSCI
>pgi|109800007|ti|559292|pi|0| YAL065C pep:known chromosome:SacCer_Apr2011:I:11565:11951:-1
gene:YAL065C transcript:YAL065C description:"Putative protein of unknown function\x3b has
```

```
homology to FLO1\x3b possible pseudogene"
MNSATSETTTNTGAAETTTSTGAAETKTVVTSSISRFNHAETQTASATDVIGHSSSVVSVSETGNTKSLITSGLSTMSQ
QPRSTPASSIIGSSTASLEISTYVGIANGLLTNNGISVFISTVLLAIVW
```

3.2 Program options

I order to see program options type:

```
./bin/SplitFasta -h
```

Expected output:

```
Usage: ./program [options]



  ______        _        _____  _
 |  ____|      | |      |  __ \| |
 | |__ __ _ ___| |_ __ _| |__) | |_   _ ___
 |  __/ _` / __| __/ _` |  ___/| | | | / __|
 | | | (_| \__ \ || (_| | |    | | |_| \__ \
 |_|  \__,_|___/\__\__,_|_|    |_|\__,_|___/

                            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
  -t [ --taxid ] arg       taxid
  -o [ --output-file ] arg output file
  -l [ --number ] arg      The number of files.
```

3.3 Example

A minimal example demonstrating the usage of SplitFasta program:

```
./bin/SplitFasta -i demo/Saccharomyces\_cerevisiae.fa -o Split -l 4

Result is a set of files located in you local directory:

Split.1
Split.2
Split.3
Split.4
```

If "-o" not specified then default name "fasta" is assigned to each file.

4 GetRandFasta

The program demonstrates how the library can be used to extract random fasta records from a given source file.

4.1 Input files

The requires a simple (multi) fasta formated file an example of which can be found in `./fastaplus-xxx/demo` directory and should look like this:

`Saccharomyces_cerevisiae.fa`:

```
>pgi|109800001|ti|559292|pi|0| YAL069W pep:known chromosome:SacCer_Apr2011:I:335:649:1
gene:YAL069W transcript:YAL069W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MIVNNTHVLTLPLYTTTTCHTHPHLYTDFTYAHGCYSIYHLKLTLLSDSTSLHGPSLTESVPNALTSLCTALASAVYTL
CHLPITPIIIHILISISHSAVPNIV
>pgi|109800002|ti|559292|pi|0| YAL068W-A pep:known chromosome:SacCer_Apr2011:I:538:792:1
gene:YAL068W-A transcript:YAL068W-A description:"Dubious open reading frame unlikely to
encode a protein\x3b identified by gene-trapping, microarray-based expression analysis, and
genome-wide homology searching"
MHGTCLSGLYPVPFTHNAHHYPHFDIYISFGGPKYCITALNTYVIPLLHHILTTPFIYTYVNITEKSPQKSPKHKNILL
FNNNT
>pgi|109800003|ti|559292|pi|0| YAL068C pep:known chromosome:SacCer_Apr2011:I:1807:2169:-1
gene:YAL068C transcript:YAL068C description:"Protein of unknown function, member of the
seripauperin multigene family encoded mainly in subtelomeric regions"
MVKLTSIAAGVAAIAATASATTTLAQSDERVNLVELGVYVSDIRAHLAQYYMFQAAHPTETYPVEVAEAVFNYGDFTTM
LTGIAPDQVTRMITGVPWYSSRLKPAISSALSKDGIYTIAN
>pgi|109800004|ti|559292|pi|0| YAL067W-A pep:known chromosome:SacCer_Apr2011:I:2480:2707:1
gene:YAL067W-A transcript:YAL067W-A description:"Putative protein of unknown function\x3b
identified by gene-trapping, microarray-based expression analysis, and genome-wide
homology searching"
MPIIGVPRCLIKPFSVPVTFPFSVKKNIRILDLDPRTEAYCLSLNSVCFKRLPRRKYFHLLNSYNIKRVLGVVYC
>pgi|109800005|ti|559292|pi|0| YAL067C pep:known chromosome:SacCer_Apr2011:I:7235:9016:-1
gene:YAL067C transcript:YAL067C description:"Putative permease, member of the allantoate
transporter subfamily of the major facilitator superfamily\x3b mutation confers resistance
to ethionine sulfoxide"
MYSIVKEIIVDPYKRLKWGFIPVKRQVEDLPDDLNSTEIVTISNSIQSHETAENFITTTSEKDQLHFETSSYSEHKDNV
NVTRSYEYRDEADRPWWRFFDEQEYRINEKERSHNKWYSWFKQGTSFKEKKLLIKLDVLLAFYSCIAYWVKYLDTVNIN
NAYVSGMKEDLGFQGNDLVHTQVMYTVGNIIFQLPFLIYLNKLPLNYVLPSLDLCWSLLTVGAAYVNSVPHLKAIRFFI
GAFEAPSYLAYQYLFGSFYKHDEMVRRSAFYYLGQYIGILSAGGIQSAVYSSLNGVNGLEGWRWNFIIDAIVSVVVGLI
GFYSLPGDPYNCYSIFLTDDEIRLARKRLKENQTGKSDFETKVFDIKLWKTIFSDWKIYILTLWNIFCWNDSNVSSGAY
LLWLKSLKRYSIPKLNQLSMITPGLGMVYLMLTGIIADKLHSRWFAIIFTQVFNIIGNSILAAWDVAEGAKWFAFMLQC
FGWAMAPVLYSWQNDICRRDAQTRAITLVTMNIMAQSSTAWISVLVWKTEEAPRYLKGFTFTACSAFCLSIWTFVVLYF
YKRDERNNAKKNGIVLYNSKHGVEKPTSKDVETLSVSDEK
>pgi|109800006|ti|559292|pi|0| YAL066W pep:known chromosome:SacCer_Apr2011:I:10091:10399:1
gene:YAL066W transcript:YAL066W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MLSLVKRSILHSIPITRHILPIQLILVKMNHVQIRNIKLYHFISYGFMLTKLTVFLFNLFFYRLRILCRLTLLILSLPV
QIYIKEIQTKMLEKHTASDTSCI
>pgi|109800007|ti|559292|pi|0| YAL065C pep:known chromosome:SacCer_Apr2011:I:11565:11951:-1
gene:YAL065C transcript:YAL065C description:"Putative protein of unknown function\x3b has
homology to FLO1\x3b possible pseudogene"
MNSATSETTTNTGAAETTTSTGAAETKTVVTSSISRFNHAETQTASATDVIGHSSSVVSVSETGNTKSLITSGLSTMSQ
QPRSTPASSIIGSSTASLEISTYVGIANGLLTNNGISVFISTVLLAIVW
```

4.2 Program options

I order to see program options type:

```
./bin/GetRandFasta -h
```

Expected output:

```
Usage: ./program [options]



  ______        _        _____  _
 |  ____|      | |      |  __ \| |
 | |__ __ _ ___| |_ __ _| |__) | |_   _ ___
 |  __/ _` / __| __/ _` |  ___/| | | | / __|
 | | | (_| \__ \ || (_| | |    | | |_| \__ \
 |_|  \__,_|___/\__\__,_|_|    |_|\__,_|___/

                            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
  -t [ --taxid ] arg       taxid
  -o [ --output-file ] arg output file
  -l [ --number ] arg      The number of random sequences to be retrieved.
```

4.3 Example

A minimal example demonstrating the usage of GetRandFasta program:

```
./bin/GetRandFasta -i demo/Saccharomyces\_cerevisiae.fa -l 3

>pgi|109800122|ti|559292|pi|0| YBL112C pep:known chromosome:SacCer_Apr2011:II:2582:2899:-1
gene:YBL112C transcript:YBL112C description:"Putative protein of unknown function\x3b
YBL112C is contained within TEL02L"
MQVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRLRDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVRE
FYGLESKKGKKGPACWMLWLQDRPVC
>pgi|109805934|ti|559292|pi|0| YOR197W pep:known chromosome:SacCer_Apr2011:XV:717086:718384:1
gene:YOR197W transcript:YOR197W description:"Putative cysteine protease similar to mammalian
caspases\x3b involved in regulation of apoptosis upon H2O2 treatment\x3b contributes to
clearance of insoluble protein aggregates during normal growth\x3b may be involved in cell
cycle progression"
MYPGSGRYTYNNAGGNNGYQRPMAPPPNQQYGQQYGQQYEQQYGQQYGQQNDQQFSQQYAPPPGPPPMAYNRPVYPPPQ
FQQEQAKAQLSNGYNNPNVNASNMYGPPQNMSLPPPQTQTIQGTDQPYQYSQCTGRRKALIIGINYIGSKNQLRGCIND
AHNIFNFLTNGYGYSSDDIVILTDDQNDLVRVPTRANMIRAMQWLVKDAQPNDSLFLHYSGHGGQTEDLDGDEEDGMDD
VIYPVDFETQGPIIDDEMHDIMVKPLQQGVRLTALFDSCHSGTVLDLPYTYSTKGIIKEPNIWKDVGQDGLQAAISYAT
GNRAALIGSLGSIFKTVKGGMGNNVDRERVRQIKFSAADVVMLSGSKDNQTSADAVEDGQNTGAMSHAFIKVMTLQPQQ
SYLSLLQNMRKELAGKYSQKPQLSSSHPIDVNLQFIM
>pgi|109801328|ti|559292|pi|0| YDR275W pep:known chromosome:SacCer_Apr2011:IV:1012252:1012959:1
gene:YDR275W transcript:YDR275W description:"Protein of unknown function, ORF exhibits genomic
```

```
organization compatible with a translational readthrough-dependent mode of expression"
MFFFPKLRKLIGSTVIDHDTKNSSGKEEIMSNSRLALVIINHAFDKVLSLTWHCGILSEIRSGLMLMFGIFQLMCSLGV
IVLLLPIIILDAIDLFLYMCRLLDYGCKLFHYNRSSLPVADGKEKTSGPISGKEEIVIDEEIINMLNESSESLINHTTA
GLEYDISSGSVNKSRRLNSTSTVTFVKQNKLVNERREDAYYEEEDDDFLSNPNYDKISLIEKSFTSRFEVACEQKAA
```

5 FilterFasta

The program demonstrates how the library can be used to clean fasta records from dubious characters, low complexity regions and repetitive segments.

5.1 Input files

The requires a simple (multi) fasta formated file an example of which can be found in `./fastaplus-xxx/demo` directory and should look like this:

`Saccharomyces_cerevisiae.fa`:

```
>pgi|109800001|ti|559292|pi|0| YAL069W pep:known chromosome:SacCer_Apr2011:I:335:649:1
gene:YAL069W transcript:YAL069W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MIVNNTHVLTLPLYTTTTCHTHPHLYTDFTYAHGCYSIYHLKLTLLSDSTSLHGPSLTESVPNALTSLCTALASAVYTL
CHLPITPIIIHILISISHSAVPNIV
>pgi|109800002|ti|559292|pi|0| YAL068W-A pep:known chromosome:SacCer_Apr2011:I:538:792:1
gene:YAL068W-A transcript:YAL068W-A description:"Dubious open reading frame unlikely to
encode a protein\x3b identified by gene-trapping, microarray-based expression analysis, and
genome-wide homology searching"
MHGTCLSGLYPVPFTHNAHHYPHFDIYISFGGPKYCITALNTYVIPLLHHILTTPFIYTYVNITEKSPQKSPKHKNILL
FNNNT
>pgi|109800003|ti|559292|pi|0| YAL068C pep:known chromosome:SacCer_Apr2011:I:1807:2169:-1
gene:YAL068C transcript:YAL068C description:"Protein of unknown function, member of the
seripauperin multigene family encoded mainly in subtelomeric regions"
MVKLTSIAAGVAAIAATASATTTLAQSDERVNLVELGVYVSDIRAHLAQYYMFQAAHPTETYPVEVAEAVFNYGDFTTM
LTGIAPDQVTRMITGVPWYSSRLKPAISSALSKDGIYTIAN
>pgi|109800004|ti|559292|pi|0| YAL067W-A pep:known chromosome:SacCer_Apr2011:I:2480:2707:1
gene:YAL067W-A transcript:YAL067W-A description:"Putative protein of unknown function\x3b
identified by gene-trapping, microarray-based expression analysis, and genome-wide
homology searching"
MPIIGVPRCLIKPFSVPVTFPFSVKKNIRILDLDPRTEAYCLSLNSVCFKRLPRRKYFHLLNSYNIKRVLGVVYC
>pgi|109800005|ti|559292|pi|0| YAL067C pep:known chromosome:SacCer_Apr2011:I:7235:9016:-1
gene:YAL067C transcript:YAL067C description:"Putative permease, member of the allantoate
transporter subfamily of the major facilitator superfamily\x3b mutation confers resistance
to ethionine sulfoxide"
MYSIVKEIIVDPYKRLKWGFIPVKRQVEDLPDDLNSTEIVTISNSIQSHETAENFITTTSEKDQLHFETSSYSEHKDNV
NVTRSYEYRDEADRPWWRFFDEQEYRINEKERSHNKWYSWFKQGTSFKEKKLLIKLDVLLAFYSCIAYWVKYLDTVNIN
NAYVSGMKEDLGFQGNDLVHTQVMYTVGNIIFQLPFLIYLNKLPLNYVLPSLDLCWSLLTVGAAYVNSVPHLKAIRFFI
GAFEAPSYLAYQYLFGSFYKHDEMVRRSAFYYLGQYIGILSAGGIQSAVYSSLNGVNGLEGWRWNFIIDAIVSVVVGLI
GFYSLPGDPYNCYSIFLTDDEIRLARKRLKENQTGKSDFETKVFDIKLWKTIFSDWKIYILTLWNIFCWNDSNVSSGAY
LLWLKSLKRYSIPKLNQLSMITPGLGMVYLMLTGIIADKLHSRWFAIIFTQVFNIIGNSILAAWDVAEGAKWFAFMLQC
FGWAMAPVLYSWQNDICRRDAQTRAITLVTMNIMAQSSTAWISVLVWKTEEAPRYLKGFTFTACSAFCLSIWTFVVLYF
YKRDERNNAKKNGIVLYNSKHGVEKPTSKDVETLSVSDEK
>pgi|109800006|ti|559292|pi|0| YAL066W pep:known chromosome:SacCer_Apr2011:I:10091:10399:1
gene:YAL066W transcript:YAL066W description:"Dubious open reading frame unlikely to encode
a protein, based on available experimental and comparative sequence data"
MLSLVKRSILHSIPITRHILPIQLILVKMNHVQIRNIKLYHFISYGFMLTKLTVFLFNLFFYRLRILCRLTLLILSLPV
QIYIKEIQTKMLEKHTASDTSCI
>pgi|109800007|ti|559292|pi|0| YAL065C pep:known chromosome:SacCer_Apr2011:I:11565:11951:-1
gene:YAL065C transcript:YAL065C description:"Putative protein of unknown function\x3b has
homology to FLO1\x3b possible pseudogene"
MNSATSETTTNTGAAETTTSTGAAETKTVVTSSISRFNHAETQTASATDVIGHSSSVVSVSETGNTKSLITSGLSTMSQ
QPRSTPASSIIGSSTASLEISTYVGIANGLLTNNGISVFISTVLLAIVW
```

5.2 Program options

I order to see program options type:

```
./bin/FilterFasta -h
```

Expected output:

```
Usage: ./program [options]



  ______        _        _____  _
 |  ____|      | |      |  __ \| |
 | |__ __ _ ___| |_ __ _| |__) | |_   _ ___
 |  __/ _` / __| __/ _` |  ___/| | | | / __|
 | | | (_| \__ \ || (_| | |    | | |_| \__ \
 |_|  \__,_|___/\__\__,_|_|    |_|\__,_|___/

                            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
  -t [ --taxid ] arg             taxid
  -o [ --output-file ] arg       output file
  -W [ --window ] arg            SEG window size.
  -H [ --hicut ] arg             High complexity cutoff.
  -L [ --locut ] arg             Low complexity cutoff.
  -T [ --maxtrim ] arg           Maximum trimming of raw segment.
  -X [ --maxxs ] arg             Maximum number of xxx characters.
  -P [ --pam ] arg               PAM  matrix to use: PAM60/PAM120/PAM250.
  -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.

Parameters, unless specified are set to their default values!!
```

5.3 Example

A minimal example demonstrating the usage of FilterFasta program:

```
./bin/FastaPlusTest -i demo/Saccharomyces\_cerevisiae.fa

>si|00000000000000000000000013307/0|ti|1|ss|0 pgi|109803307|ti|559292|pi|0| YJL172W pep:known
chromosome:SacCer_Apr2011:X:97737:99467:1 gene:YJL172W transcript:YJL172W description:
"Vacuolar carboxypeptidase yscS\x3b expression is induced under low-nitrogen conditions"
RAW:
MIALPVEKAPRKSLWQRHRAFISGIVALIIIGTFFLTSGLHPAPPHEAKRPHHGKGPMHSPKCEKIEPLSPSFKHSVDT
ILHDPAFRNSSIEKLSNAVRIPTVVQDKNPNPADDPDFYKHFYELHDYFEKTFPNIHKHLKLEKVNELGLLYTWEGSDP
DLKPLLLMAHQDVVPVNNETLSSWKFPPFSGHYDPETDFVWGRGSNDCKNLLIAEFEAIEQLLIDGFKPNRTIVMSLGF
```

```
DEEASGTLGAASLASFLHERYGDDGIYSIIDEGEGIMEVDKDVFVATPINAEKGYVDFEVSILGHGGHSSVPPDHTTIG
IASELITEFEANPFDYEFEFDNPIYGLLTCAAEHSKSLSKDVKKTILGAPFCPRRKDKLVEYISNQSHLRSLIRTTQAV
DIINGGVKANALPETTRFLINHRINLHSSVAEVFERNIEYAKKIAEKYGYGLSKNGDDYIIPETELGHIDITLLRELEP
APLSPSSGPVWDILAGTIQDVFENGVLQNNEEFYVTTGLFSGNTDTKYYWNLSKNIYRFVGSIIDIDLLKTLHSVNEHV
DVPGHLSAIAFVYEYIVNVNEYA
SEG:
MIALPVEKAPRKSLWQRHRAFISGIVALIIIGTFFLTSGLHPAPPHEAKRPHHGKGPMHSPKCEKIEPLSPSFKHSVDT
ILHDPAFRNSSIEKLSNAVRIPTVVQDKNPNPADDPDFYKHFYELHDYFEKTFPNIHKHLKLEKVNELGLLYTWEGSDP
DLKPLLLMAHQDVVPVNNETLSSWKFPPFSGHYDPETDFVWGRGSNDCKNLLIAEFEAIEQLLIDGFKPNRTIVMSLGF
DEEXXXXXXXXXXXXFLHERYGDDGIYSIIDEGEGIMEVDKDVFVATPINAEKGYVDFEVSILGHGGHSSVPPDHTTIG
IASELITEFEANPFDYEFEFDNPIYGLLTCAAEHSKSLSKDVKKTILGAPFCPRRKDKLVEYISNQSHLRSLIRTTQAV
DIINGGVKANALPETTRFLINHRINLHSSVAEVFERNIEYAKKIAEKYGYGLSKNGDDYIIPETELGHIDITLLRELEP
APLSPSSGPVWDILAGTIQDVFENGVLQNNEEFYVTTGLFSGNTDTKYYWNLSKNIYRFVGSIIDIDLLKTLHSVNEHV
DVPGHLSAIAFVYEYIVNVNEYA
XNU:
MIALPVEKAPRKSLWQRHRAFISGIVALIIIGTFFLTSGLHPAPPHEAKRPHHGKGPMHSPKCEKIEPLSPSFKHSVDT
ILHDPAFRNSSIEKLSNAVRIPTVVQDKNPNPADDPDFYKHFYELHDYFEKTFPNIHKHLKLEKVNELGLLYTWEGSDP
DLKPLLLMAHQDVVPVNNETLSSWKFPPFSGHYDPETDFVWGRGSNDCKNLLIAEFEAIEQLLIDGFKPNRTIVMSLGF
DEEASGTLGAASLASFLHERYGDDGIYSIIDEGEGIMEVDKDVFVATPINAEKGYVDFEVSILGHGGHSSVPPDHTTIG
IASELITEFEANPXXXXXXXXNPIYGLLTCAAEHSKSLSKDVKKTILGAPFCPRRKDKLVEYISNQSHLRSLIRTTQAV
DIINGGVKANALPETTRFLINHRINLHSSVAEVFERNIEYAKKIAEKYGYGLSKNGDDYIIPETELGHIDITLLRELEP
APLSPSSGPVWDILAGTIQDVFENGVLQNNEEFYVTTGLFSGNTDTKYYWNLSKNIYRFVGSIIDIDLLKTLHSVNEHV
DVPGHLSAIAFVYEYIVNVNEYA
SEG+XNU:
MIALPVEKAPRKSLWQRHRAFISGIVALIIIGTFFLTSGLHPAPPHEAKRPHHGKGPMHSPKCEKIEPLSPSFKHSVDT
ILHDPAFRNSSIEKLSNAVRIPTVVQDKNPNPADDPDFYKHFYELHDYFEKTFPNIHKHLKLEKVNELGLLYTWEGSDP
DLKPLLLMAHQDVVPVNNETLSSWKFPPFSGHYDPETDFVWGRGSNDCKNLLIAEFEAIEQLLIDGFKPNRTIVMSLGF
DEEXXXXXXXXXXXXFLHERYGDDGIYSIIDEGEGIMEVDKDVFVATPINAEKGYVDFEVSILGHGGHSSVPPDHTTIG
IASELITEFEANPXXXXXXXXNPIYGLLTCAAEHSKSLSKDVKKTILGAPFCPRRKDKLVEYISNQSHLRSLIRTTQAV
DIINGGVKANALPETTRFLINHRINLHSSVAEVFERNIEYAKKIAEKYGYGLSKNGDDYIIPETELGHIDITLLRELEP
APLSPSSGPVWDILAGTIQDVFENGVLQNNEEFYVTTGLFSGNTDTKYYWNLSKNIYRFVGSIIDIDLLKTLHSVNEHV
DVPGHLSAIAFVYEYIVNVNEYA
```

6 Acknowledgement

1. Wootton, J.C., Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry 17: 149-163.
2. Jean Michel Claverie & David J. States (1993) Computers and Chemistry 17: 191-201.

7 Future work

1. Implement DNA sequence filters.