Perl for Biologists

Session 11

May 13, 2015

Object Oriented Programming and BioPERL (2)

Qi Sun

Session 10 Exercises

Exercise 1. Translate all DNA sequences in a Fasta file

```
#!usr/local/bin/perl
use strict;
use warnings;

use Bio::SeqIO;

my $in  = Bio::SeqIO->new(-file => "/home/jarekp/perl_10/yeast_orf.fasta" ,
                                        -format => 'Fasta');
my $out = Bio::SeqIO->new(-file => ">yeast_pep.fasta" ,
                                        -format => 'Fasta');

while ( my $seqobj = $in->next_seq() )
{
        my $proteinSeqObj = $seqobj->translate();
        $proteinSeqObj->display_id($seqobj->display_id . "_pep");
        $proteinSeqObj->desc("");
        $out->write_seq($proteinSeqObj);
};
```

Session 10 Exercises

Exercise 2. Make a fasta file with 10 random sequences

```
#!/usr/local/bin/perl
use strict;
use warnings;

use String::Random;
use Bio::SeqIO;

my $out = Bio::SeqIO->new(-file => ">random_dna.fasta" ,
                                  -format => 'Fasta');

my $RandomSeq =  String::Random->new();

for (my $i=0; $i<10; $i++)
{
        my $seqstr= $RandomSeq->randregex('[ACGT]{1000}');

        my $seqObject = Bio::Seq->new (-seq => $seqstr,
                                       -display_id => "seq$i",
                                       -alphabet => "dna");
        $out->write_seq($seqObject);
}
```

Review of Session 10

Bio::Seq object

A Constructor:

```
my $seqObject = Bio::Seq->new (-seq => “AAAACCCCTTGGGAAGC”,
                               -display_id => “myseq1”,
                               -desc => “This is an example.",
                               -alphabet => "dna");
```

Methods

```
$seqObject -> revcom() -> translate(-frame=>0);
```

Alternative ways to create the sequence objects

1. From network database (e.g. NCBI Genbank)

```
use Bio::Perl;
$db = Bio::DB::GenBank->new();
$seqobj = $db->get_Seq_by_acc('X78121');
```

2. From file

```
use Bio::SeqIO;
$in = Bio::SeqIO->new(-file => "inputfile.fasta" ,
                      -format => 'Fasta');
while ( my $seqobj = $in->next_seq() )
{
    …
}
```

Other properties of Bio::Seq object

GenBank File Format

```
LOCUS       NC_000913            4639675 bp    DNA     circular BCT 04-MAR-2013
DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION   NC_000913
VERSION     NC_000913.2  GI:49175990
DBLINK      Project: 57779
            BioProject: PRJNA57779
KEYWORDS    .
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
FEATURES             Location/Qualifiers
     source          1..4639675
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     gene            190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001; JW4367"
                     /db_xref="EcoGene:EG11277"
                     /db_xref="GeneID:944742"
     CDS             190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001; JW4367"
                     /function="leader; Amino acid biosynthesis: Threonine"
                     /function="1.5.1.8 metabolism; building block
                     biosynthesis; amino acids; threonine"
                     /GO_process="GO:0009088 - threonine biosynthetic process"
```

Other Bio::Seq properties: Seq Features GFF3 File Format

Chr1	TAIR10	chromosome	1	30427671	.	.	.	ID=Chr1;Name=Chr1
Chr1	TAIR10	gene	3631	5899	.	+	.	ID=AT1G01010;Note=prote
Chr1	TAIR10	mRNA	3631	5899	.	+	.	ID=AT1G01010.1;Parent=AT
Chr1	TAIR10	protein	3760	5630	.	+	.	ID=AT1G01010.1-Protein;N
Chr1	TAIR10	exon	3631	3913	.	+	.	Parent=AT1G01010.1
Chr1	TAIR10	five_prime_UTR		3631	3759	.	+	. Parent=AT1G
Chr1	TAIR10	CDS	3760	3913	.	+	0	Parent=AT1G01010.1,AT1G
Chr1	TAIR10	exon	3996	4276	.	+	.	Parent=AT1G01010.1
Chr1	TAIR10	CDS	3996	4276	.	+	2	Parent=AT1G01010.1,AT1G
Chr1	TAIR10	exon	4486	4605	.	+	.	Parent=AT1G01010.1
Chr1	TAIR10	CDS	4486	4605	.	+	0	Parent=AT1G01010.1,AT1G
Chr1	TAIR10	exon	4706	5095	.	+	.	Parent=AT1G01010.1
Chr1	TAIR10	CDS	4706	5095	.	+	0	Parent=AT1G01010.1,AT1G
Chr1	TAIR10	exon	5174	5326	.	+	.	Parent=AT1G01010.1
Chr1	TAIR10	CDS	5174	5326	.	+	0	Parent=AT1G01010.1,AT1G
Chr1	TAIR10	exon	5439	5899	.	+	.	Parent=AT1G01010.1
Chr1	TAIR10	CDS	5439	5630	.	+	0	Parent=AT1G01010.1,AT1G
Chr1	TAIR10	three_prime_UTR		5631	5899	.	+	. Parent=AT1G
Chr1	TAIR10	gene	5928	8737	.	-	.	ID=AT1G01020;Note=prote
Chr1	TAIR10	mRNA	5928	8737	.	-	.	ID=AT1G01020.1;Parent=AT
Chr1	TAIR10	protein	6915	8666	.	-	.	ID=AT1G01020.1-Protein;N
Chr1	TAIR10	five_prime_UTR		8667	8737	.	-	. Parent=AT1G
Chr1	TAIR10	CDS	8571	8666	.	-	0	Parent=AT1G01020.1,AT1G
Chr1	TAIR10	exon	8571	8737	.	-	.	Parent=AT1G01020.1
Chr1	TAIR10	CDS	8417	8464	.	-	0	Parent=AT1G01020.1,AT1G
Chr1	TAIR10	exon	8417	8464	.	-	.	Parent=AT1G01020.1
Chr1	TAIR10	CDS	8236	8325	.	-	0	Parent=AT1G01020.1,AT1G

Other Bio::Seq properties: Seq Features GFF3 File Format

Chr1	TAIR10	chromosome	1	30427671	.	.	.	ID=Chr1;Name=Chr1
Chr1	TAIR10	gene	3631	5899	.	+	.	ID=AT1G01010;Note=prote
Chr1	TAIR10	mRNA	3631	5899	.	+	.	ID=AT1G01010.1;Parent=AT
Chr1								01010.1-Protein;N
C								01010.1
								Parent=AT1G
								010.1,AT1G
								010.1
								010.1,AT1G
								010.1
								010.1,AT1G
								010.1
								010.1,AT1G
								010.1
								010.1,AT1G
								010.1
								010.1,AT1G
								Parent=AT1G
								Note=prote
								1;Parent=AT
								0.1-Protein;N
Ch								Parent=AT1G
Chr1	TAIR10	CDS	8571	8666	.	-	0	Parent=AT1G01020.1,AT1G
Chr1	TAIR10	exon	8571	8737	.	-	.	Parent=AT1G01020.1
Chr1	TAIR10	CDS	8417	8464	.	-	0	Parent=AT1G01020.1,AT1G
Chr1	TAIR10	exon	8417	8464	.	-	.	Parent=AT1G01020.1
Chr1	TAIR10	CDS	8236	8325	.	-	0	Parent=AT1G01020.1,AT1G

```
open (IN, “tair10.gff3”) || die “Can not open GFF3 file!\n ”;

while (<IN>)
{
        my @data = split “\t”;
        …
}
```

Retrieve seq features from a Bio:Seq object constructed from NCBI Genbank

script1.pl

```
#!/usr/local/bin/perl
use strict;
use warnings;
use Bio::Perl;

my $db = Bio::DB::GenBank->new();
my $seqobj = $db->get_Seq_by_acc('NC_000913');

$,="\t";
my $count =0;
for my $feat_object ($seqobj->get_SeqFeatures) {
   if ($feat_object->primary_tag eq "gene") {
         $count ++;
         print $feat_object->get_tag_values('gene'),
               $feat_object->start(),
               $feat_object->end(),
               $feat_object->strand(),"\n";
   }
}
print "Total number of genes: $count\n";
```

Run Sequence Analysis tools

1. <u>Using BioPERL wrapper: Bio::Tools::Run</u>

ClusterW
MUSCLE
BLAST
…
Primer3
…

2. <u>Using system calling</u>

```
system ("primer3_core < inputFile");
```

Or

```
my $stdout = ` primer3_core < inputFile `;
```

Using Bio::Tools::Run::Primer3

script2.pl

```
#!/usr/local/bin/perl
use strict;
use warnings;
use Bio::DB::GenBank;
use Bio::Tools::Run::Primer3;

my $db = Bio::DB::GenBank->new();
my $seqobj = $db->get_Seq_by_acc('NM_001126114');

my $primer3 = Bio::Tools::Run::Primer3->new(
        -seq => $seqobj,
    -outfile => "temp.out",
    -path => "/programs/primer3-2.3.5/src/primer3_core");

$primer3->add_targets(
        "PRIMER_MIN_TM"=>56,
        "PRIMER_MAX_TM"=>90,
        "PRIMER_MIN_SIZE"=>18,
        "PRIMER_MAX_SIZE"=>21);

my $results = $primer3->run;
print "There were ", $results->number_of_results, " primers\n";
```

Bio::Tools::Run::Primer3 does not work with latest version of Primer3

Parameter name is changed after Primer3 2.0

Boulder data interchange format

```
SEQUENCE_ID=example
SEQUENCE=GTAGTCAGTAGACNATGACNACTGACGATGCAGACNAC
ACACACACACACAGCACACAGGTATTAGTGGGCCATTCGATCCCGACC
CAAATCGATAGCTACGATGACG
SEQUENCE_TARGET=37,21
PRIMER_TASK=pick_detection_primers
PRIMER_PICK_LEFT_PRIMER=1
PRIMER_PICK_INTERNAL_OLIGO=1
PRIMER_PICK_RIGHT_PRIMER=1
PRIMER_OPT_SIZE=18
PRIMER_MIN_SIZE=15
PRIMER_MAX_SIZE=21
PRIMER_MAX_NS_ACCEPTED=1
PRIMER_PRODUCT_SIZE_RANGE=75-100
P3_FILE_FLAG=1
SEQUENCE_INTERNAL_EXCLUDED_REGION=37,21
PRIMER_EXPLAIN_FLAG=1
=
```

Tag name is changed to **SEQUENCE_TEMPLATE** In latest version.

Using Bio::Tools::Run::Primer3

script3.pl

```
#!/usr/local/bin/perl
use strict;
use warnings;
use Bio::DB::GenBank;
my $PRIMER_MIN_TM=56;
my $PRIMER_MAX_TM=90;
my $PRIMER_MIN_SIZE=15;
my $PRIMER_MAX_SIZE=21;

my $db = Bio::DB::GenBank->new();
my $seqobj = $db->get_Seq_by_acc('NM_001126114');
my $seqid = $seqobj->display_id();
my $seqstr = $seqobj->seq();

open OUT, ">temp.input";
print OUT <<EOF;
SEQUENCE_ID=$seqid
SEQUENCE_TEMPLATE=$seqstr;
PRIMER_MIN_TM=$PRIMER_MIN_TM
PRIMER_MAX_TM=$PRIMER_MAX_TM
PRIMER_MIN_SIZE=$PRIMER_MIN_SIZE
PRIMER_MAX_SIZE=$PRIMER_MAX_SIZE
PRIMER_LIBERAL_BASE=1
=
EOF
close OUT;
system "/programs/primer3-2.3.5/src/primer3_core -output=temp.output temp.input";
```

Parsing results from analysis software

Output from codeml

```
……................
……................
Model 1: NearlyNeutral (2 categories)


TREE # 1: ((3, 4), 2, (1, 5)); MP score: 0
lnL(ntime: 7 np: 10): -548.665307 +0.000000
  6..7    7..3    7..4    6..2    6..8    8..1    8..5
 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 1.98425 0.60769 0.54695
Note: Branch length is defined as number of nucleotide substitutions per codon (not per neucleotide site).

tree length =  0.00002

((3: 0.000005, 4: 0.000005): 0.000000, 2: 0.000005, (1: 0.000005, 5: 0.000005): 0.000000);

((CT18: 0.000005, Ty2: 0.000005): 0.000000, ch: 0.000005, (ATCC9150: 0.000005, LT2: 0.000005): 0.000000);

Detailed output identifying parameters

kappa (ts/tv) =  1.98425


dN/dS for site classes (K=2)

p:   0.60769  0.39231
w:   0.54695  1.00000

dN & dS for each branch

 branch       t      S      N   dN/dS     dN     dS  S*dS  N*dN

   6..7    0.000  116.9  291.1  0.7247  0.0000  0.0000   0.0   0.0
   7..3    0.000  116.9  291.1  0.7247  0.0000  0.0000   0.0   0.0
   7..4    0.000  116.9  291.1  0.7247  0.0000  0.0000   0.0   0.0
   6..2    0.000  116.9  291.1  0.7247  0.0000  0.0000   0.0   0.0
   6..8    0.000  116.9  291.1  0.7247  0.0000  0.0000   0.0   0.0
   8..1    0.000  116.9  291.1  0.7247  0.0000  0.0000   0.0   0.0
   8..5    0.000  116.9  291.1  0.7247  0.0000  0.0000   0.0   0.0

……................
……................
```

Parsing results from analysis software

Output from codeml

……
……..
Model 1: NearlyNeutral (2 categories)

TR

6..2
6..8 0.000 116.9 291.1 0.7247 0.0000 0.0000 0.0 0.0
8..1 0.000 116.9 291.1 0.7247 0.0000 0.0000 0.0 0.0
8..5 0.000 116.9 291.1 0.7247 0.0000 0.0000 0.0 0.0

……..
……

PAML Parser

```
use Bio::Tools::Phylo::PAML;
my $parser = Bio::Tools::Phylo::PAML->new(
                -file => "./output.mlc",
                -dir  => "./",
                -ctlf => "./codeml.ctl");

while(my $result = $parser->next_result) {
    # do something with the results from this dataset
        ...
}
```

Parse Blast Results

```
blastall -p blastp -i rice.fasta -d TAIR7_pep_db -o blastresults
```

Note:
Most new software starts to provide machine readable output files, e.g. NCBI BLAST

-m 7 : XML (used by Blast2GO, et al.)
-m 8 : tab delimited text file (used by OrthoMCL, et al.)

BLAST Results

```
Query= Os01g01010.1
         (702 letters)

Database: TAIR7_pep_20070320
           31,921 sequences; 13,036,889 total letters

Searching..................................................done

                                                            Score    E
Sequences producing significant alignments:                 (bits) Value

AT2G43490.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   621   0.0
AT3G59570.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   608   0.0
AT5G54780.1 | Symbols:  | RAB GTPase activator | chr5:22265922-2...   184   4e-051
AT4G27100.2 | Symbols:  | RAB GTPase activator | chr4:13595851-1...   183   6e-051
AT4G27100.1 | Symbols:  | RAB GTPase activator | chr4:13595851-1...   182   9e-051
AT2G20440.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   175   4e-048
AT4G28550.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   170   2e-046
AT5G41940.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   136   6e-034
AT5G53570.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   134   4e-033
AT5G24390.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   130   5e-032
……................
……................
```

BLAST Results

Query Object
Query name
Query length

```
Query= Os01g01010.1
       (702 letters)

Database: TAIR7_pep_20070320
           31,921 sequences; 13,036,889 total letters

Searching..................................................done

                                                            Score    E
Sequences producing significant alignments:                 (bits) Value

AT2G43490.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   621   0.0
AT3G59570.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   608   0.0
AT5G54780.1 | Symbols:  | RAB GTPase activator | chr5:22265922-2...   184   4e-051
AT4G27100.2 | Symbols:  | RAB GTPase activator | chr4:13595851-1...   183   6e-051
AT4G27100.1 | Symbols:  | RAB GTPase activator | chr4:13595851-1...   182   9e-051
AT2G20440.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   175   4e-048
AT4G28550.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   170   2e-046
AT5G41940.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   136   6e-034
AT5G53570.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   134   4e-033
AT5G24390.1 | Symbols:  | RabGAP/TBC domain-containing protein |...   130   5e-032
……................
……................
```

```
>AT4G27100.2 RAB GTPase activator
        Length = 433

 Score =  183 bits (464), Expect = 6e-051,   Method: Compositional matrix adjust.
 Identities = 91/188 (48%), Positives = 122/188 (64%), Gaps = 10/188 (5%)

Query: 370 GTKSNSVVASKD---------RVSEWLWTLHRIVVDVVRTDSHLDFYGESRNMARMSDIL 420
           GT SN  V  K+         ++ +WL TLH+I +DV RTD  L FY +  N++++ DIL
Sbjct: 144 GTNSNGSVFFKELTSRGPLDKKIIQWLLTLHQIGLDVNRTDRALVFYEKKENLSKLWDIL 203

Query: 421 AVYAWVDPSTGYCQGMSDLLSPFVVLYEDDADAFWCFEMLLRRMRENFQMEG-PTGVMKQ 479
           +VYAW+D   GYCQGMSDL SP ++L ED+ADAFWCFE L+RR+R NF+  G   GV  Q
Sbjct: 204 SVYAWIDNDVGYCQGMSDLCSPMIILLEDEADAFWCFERLMRRLRGNFRSTGRSVGVEAQ 263

Query: 480 LQALWKIMEITDVELFEHLSTIGAESLHFAFRMLLVLFRRELSFEESLSMWEMMWAADFN 539
           L  L  I ++ D +L +HL  +G     FA RML+V FRRE SF +SL +WEMMWA +++
Sbjct: 264 LTHLSSITQVVDPKLHQHLDKLGGGDYLFAIRMLMVQFRREFSFCDSLYLWEMMWALEYD 323

Query: 540 EDVILHLE 547
            D+    E
Sbjct: 324 PDLFYVYE 331


 Score = 65.1 bits (157), Expect = 5e-011,   Method: Compositional matrix adjust.
 Identities = 42/96 (43%), Positives = 54/96 (56%), Gaps = 3/96 (3%)

Query: 55  VKGSKMLKPEKWHTCFDNDGKV-IGFRKALKFIVLGGVDPTIRAEVWEFLLGCYALSSTS 113
           +K  K L   KW   F  +G + IG  K L+ I  GG+ P+IR EVWEFLLGCY   ST
Sbjct: 29  IKPGKTLSVRKWQAVFVQEGSLHIG--KTLRRIRRGGIHPSIRGEVWEFLLGCYDPMSTF 86

Query: 114 EYRRKLRAVRREKYQILVRQCQSMHPSIGTGELAYA 149
           E R ++R  RR +Y     +C+ M P IG+G    A
Sbjct: 87  EEREQIRQRRRLQYASWKEECKQMFPVIGSGRFTTA 122
```

Hit

HSP 1

HSP 2

```
>AT4G27100.2 RAB GTPase activator
        Length = 433

 Score =  183 bits (464), Expect = 6e-051,   Method:                    adjust.
 Identities = 91/188 (48%), Positives = 122/188 (64%

Query: 370 GTKSNSVVASKD---------RVSEWLWTLHRIV
           GT SN  V  K+         ++ +WL TLH+I
Sbjct: 144 GTNSNGSVFFKELTSRGPLDKKIIQWLLTLHQIC

Query: 421 AVYAWVDPSTGYCQGMSDLLSPFVVLYEDDADAF
           +VYAW+D   GYCQGMSDL SP ++L ED+ADAF
Sbjct: 204 SVYAWIDNDVGYCQGMSDLCSPMIILLEDEADAF

Query: 480 LQALWKIMEITDVELFEHLSTIGAESLHFAFRML
           L  L  I ++ D +L +HL  +G     FA RML
Sbjct: 264 LTHLSSITQVVDPKLHQHLDKLGGGDYLFAIRML

Query: 540 EDVILHLE 547
            D+    E
Sbjct: 324 PDLFYVYE 331


 Score = 65.1 bits (157), Expect = 5e-011,   Method: Compositional matrix adjust.
 Identities = 42/96 (43%), Positives = 54/96 (56%), Gaps = 3/96 (3%)

Query: 55  VKGSKMLKPEKWHTCFDNDGKV-IGFRKALKFIVLGGVDPTIRAEVWEFLLGCYALSSTS 113
           +K  K L   KW   F  +G + IG  K L+ I  GG+ P+IR EVWEFLLGCY   ST
Sbjct: 29  IKPGKTLSVRKWQAVFVQEGSLHIG--KTLRRIRRGGIHPSIRGEVWEFLLGCYDPMSTF 86

Query: 114 EYRRKLRAVRREKYQILVRQCQSMHPSIGTGELAYA 149
           E R ++R  RR +Y     +C+ M P IG+G    A
Sbjct: 87  EEREQIRQRRRLQYASWKEECKQMFPVIGSGRFTTA 122
```

Each hit object:
Hit name
Hit length
Hsps

Each HSP object:
Query: start - end - strand
Hit: start - end – strand
Bit score
E-value
Identities
Positives
Alignment length
Gaps
Query sequence
Hit sequence

Hit

HSP 1

HSP 2

BLAST Parser

script4.pl

```
#!/usr/local/bin/perl
use Bio::SearchIO;
($infile, $outfile) = @ARGV;

open OUT, ">$outfile";
$,="\t";

$searchio = Bio::SearchIO->new(-format => 'blast',
                               -file   => $infile);
while ($result = $searchio->next_result)
{

        # Get info about the entire report
        $query_name = $result->query_name;
        $query_length = $result->query_length;


        # get info about the first hit
        while ($hit = $result->next_hit)
        {
                $hit_name = $hit->name;
                $hit_length = $hit->length;

                # get info about the first hsp of the first hit
                while ($hsp = $hit->next_hsp)
                {
                        $rank = $hsp->rank;
                        $num_conserved = $hsp->num_conserved ;
```

Loop1: Query

Loop2: Hit

Loop3: HSP

BLAST Parser

script4.pl

```
        while ($hit = $result->next_hit)
        {
                $hit_name = $hit->name;
                $hit_length = $hit->length;

                # get info about the first hsp of the first hit
                while ($hsp = $hit->next_hsp){
                        $rank = $hsp->rank;
                        $num_conserved = $hsp->num_conserved ;
                        $num_identical= $hsp->num_identical ;
                        $hsp_length= $hsp->hsp_length ;
                        $bits= $hsp->bits ;
                        $evalue = $hsp->evalue ;
                        $hsp_qstart = $hsp->query->start;
                        $hsp_qend = $hsp->query->end;
                        $query_strand = $hsp->query->strand;
                        $hsp_hstart = $hsp->hit->start;
                        $hsp_hend = $hsp->hit->end;
                        $hit_strand = $hsp->hit->strand;
                        $query_string = $hsp->query_string ;
                        $hit_string = $hsp->hit_string ;
                        $homology_string = $hsp->homology_string ;
                        print OUT 1, $query_name, $hit_name, $query_length, $hit_length, $rank, $num_identical,
$num_conserved, $hsp_length, $bits, $evalue, $hsp_qstart, $hsp_qend, $query_strand, $hsp_hstart, $hsp_hend, $hit_strand,
$query_string, $hit_string, $homology_string, "", "";
                print  OUT "\n";
                }
        }
}
```

HSP

Hit

Query

Parsed results from BLAST

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	1	Os01g010:	AT2G4349	702	756	1	345	452	730	621	0	12	685	0	3	689	0	TAGDYIKWMCXXXXXX	SGGEGKQWSCGKA	+ G+ +W C	+LQ+ VGSLVRD+	+P
2	1	Os01g010:	AT3G5957	702	720	1	334	435	687	608	0	12	685	0	3	653	0	TAGDYIKWMCXXXXXX	SAGEGKKW-----TRF	+AG+ KW	+LQ+ V SLVRD+	+PC
3	1	Os01g010:	AT5G5478	702	432	1	109	157	300	184	4.00E-51	382	680	0	163	414	0	RVSEWLWTLHRIVVDV\	KVIQWLLTLHQIGLD	+V +WL TLH+I +DV	RTD L FY +	N+++
4	1	Os01g010:	AT5G5478	702	432	2	40	51	95	64.7	7.00E-11	55	149	0	29	122	0	VKGSKMLKPEKWHTCFI	IKPGKTLSVRKWQA	+K K L KW F +G +	K L I GG+	P+IF
5	1	Os01g010:	AT4G2710	702	433	1	91	122	188	183	6.00E-51	370	547	0	144	331	0	GTKSNSVVASKD--------	GTNSNGSVFFKELTS	GT SN V K+	++ +WL TLH+I	+DV F
6	1	Os01g010:	AT4G2710	702	433	2	42	54	96	65.1	5.00E-11	55	149	0	29	122	0	VKGSKMLKPEKWHTCFI	IKPGKTLSVRKWQA	+K K L KW F +G +	IG K L+ I GG+	P-
7	1	Os01g010:	AT4G2710	702	433	3	17	30	51	37	0.035	632	680	0	365	415	0	AKNGDDDLPI--FCVAAI	GKSAEGPLPISVFLV	K+ + LPI F VA++L	+K++ E R	+DD +K
8	1	Os01g010:	AT4G2710	702	436	1	91	122	188	182	9.00E-51	370	547	0	144	331	0	GTKSNSVVASKD--------	GTNSNGSVFFKELTS	GT SN V K+	++ +WL TLH+I	+DV F
9	1	Os01g010:	AT4G2710	702	436	2	42	54	96	65.1	5.00E-11	55	149	0	29	122	0	VKGSKMLKPEKWHTCFI	IKPGKTLSVRKWQA	+K K L KW F +G +	IG K L+ I GG+	P-
10	1	Os01g010:	AT4G2710	702	436	3	17	30	51	37	0.036	632	680	0	365	415	0	AKNGDDDLPI--FCVAAI	GKSAEGPLPISVFLV	K+ + LPI F VA++L	+K++ E R	+DD +K
11	1	Os01g010:	AT2G2044	702	425	1	95	157	310	175	4.00E-48	377	685	0	154	409	0	VASKDRVSEWLWTLHRI	TVTDERVLQWMLSL	+ +RV +W+ +LH+I	+DV RTD	+L FY R
12	1	Os01g010:	AT2G2044	702	425	2	39	51	90	80.1	9.00E-16	56	145	0	37	125	0	KGSKMLKPEKWHTCFDI	RAGKTLSARRWHAA	+ K L +WH F DG +	K L+ I GG+	P+I-
13	1	Os01g010:	AT4G2855	702	424	1	79	112	168	170	2.00E-46	382	548	0	159	326	0	RVSEWLWTLHRIVVDV\	RVLQWMLVLSQIGL	RV +W+ L +I +DV	VRTD +L FY	N AR+
14	1	Os01g010:	AT4G2855	702	424	2	39	53	92	80.9	4.00E-16	56	147	0	37	127	0	KGSKMLKPEKWHTCFDI	RAGKTLSARKWHAA	+ K L KWH F DG +	+ L+ I GG+	P+I-
15	1	Os01g010:	AT5G4194	702	549	1	60	84	122	136	6.00E-34	414	535	0	345	465	0	ARMSDILAVYAWVDPST	ARLVGILEAYAVYDP	AR+ IL YA DP	GYCQGMSDLLSP	+ +
16	1	Os01g010:	AT5G4194	702	549	2	48	72	135	72.4	4.00E-13	45	176	0	76	205	0	IGDPCLNPSPVKGSKMLI	IGSPW---SLRRRKRV	IG P S + ++L+P++W+ F	+G++ G K	
17	1	Os01g010:	AT5G5357	702	550	1	59	86	122	134	4.00E-33	414	535	0	338	458	0	ARMSDILAVYAWVDPST	ARLVAILEAYAMYDF	AR+ IL YA DP	GYCQGMSDLLSP	+ +
18	1	Os01g010:	AT5G5357	702	550	2	37	53	81	72.8	3.00E-13	61	137	0	96	173	0	LKPEKWHTCFDNDGKV-	LTPHQWRSLFTPEGI	L P +W + F +GK+	+GF LK +	GVDP+
19	1	Os01g010:	AT5G2439	702	528	1	56	87	124	130	5.00E-32	412	535	0	315	437	0	NMARMSDILAVYAWVD	HAARLVAVLEAYALF	+ AR+ +L YA DP	GYCQGMSDLLSP	+
20	1	Os01g010:	AT5G2439	702	528	2	38	52	96	65.9	4.00E-11	47	138	0	46	141	0	DPCLNPSP---VKGSKML	DPHRLKSPWSRRKG	DP SP KG K L	+W CF +G++ G	L
21	1	Os01g010:	AT3G4935	702	539	1	57	88	133	127	9.00E-31	403	535	0	320	450	0	HLDFYGESRNMARMSDI	HLEPY-KIFQAARLV	HL+ Y + AR+ +L	YA DP GYCQGMSI	
22	1	Os01g010:	AT3G4935	702	539	2	30	46	76	60.8	2.00E-09	64	138	0	72	147	0	EKWHTCFDNDGKVI-GF	QQWKRFFTPDGRLF	++W F DG++ G	LK + G++P+IR	EV
23	1	Os01g010:	AT5G5259	702	338	1	57	90	153	112	8.00E-27	396	546	0	107	258	0	DVVRTDSHLDFYGESRN	DVVRTDRAFEYYEGI	DVVRTD ++Y N+	M DIL Y++ +	G
24	1	Os01g010:	AT5G5259	702	338	2	34	50	93	59.7	2.00E-09	70	159	0	19	111	0	FDNDGKVIGFRKALKFIV	LDSEGRVVESKALRE	D++G+V+ + + +	GG++ +R EVW	FLL
25	1	Os01g010:	AT4G1373	702	408	1	49	91	177	77.4	6.00E-15	389	546	0	227	403	0	TLHRIVVDVVRTDSHLDF	VLEQIERDVMRTHPI	L +I DV+RT + F+	+A+ + +IL	++A
26	1	Os01g010:	AT4G1373	702	449	1	47	87	164	76.6	1.00E-14	389	533	0	227	390	0	TLHRIVVDVVRTDSHLDF	VLEQIERDVMRTHPI	L +I DV+RT + F+	+A+ + +IL	++A
27	1	Os01g010:	AT1G0483	702	448	1	56	90	186	72.8	2.00E-13	389	555	0	223	398	0	TLHRIVVDVVRTDSHLDF	TIEQIDRDVKRTHPD	T+ +I DV RT + F+	+ AR M	+IL V+A
28	1	Os01g010:	AT2G3071	702	440	1	45	79	169	71.6	5.00E-13	390	535	0	201	368	0	LHRIVVDVVRTDSHLDF\	LRQIAVDCPRTVPD\	L +I VD RT + F+ + +	+ IL +A	P++G
29	1	Os01g010:	AT2G1924	702	840	1	32	51	109	49.3	7.00E-06	444	537	0	248	356	0	VVLYED--DADAFWCFEI	IVLSEKFMEHDAYCN	+VL E + DA+ F+ L+ +	F M	G

Sequence Alignment

Clustalw format

```
CLUSTAL W(1.81) multiple sequence alignment


seq1            VANITLSTQHYRIHRSDVEPVKEKTTDKDVFAKSITAVRNSFISLSTSLSDRFSLHLQTD
seq2            VTNITLSTQHYRIHRSDVEPVKEKTTEKDIFAKSITAVRNSFISLSTSLSDRFSLHQQTD
seq3            VTNITLSTQHYRIHRSDVEPVKEKTTEKDIFAKSITAVRNSFISLSTSLSDRFSLHQQTD
seq4            VTKITLSPQNFRIQKQET--LKEKSTEKNSLAKSILAVKNHFIELRSKLSERFISHKNTE
seq5            VTKITLSPQNFRIQKQETTLLKEKSTEKNSLAKSILAVKNHFIELRSKLSERFISHKNTE
                *::****.*::**::.:.  :***:*:*: :**** **:* **.* :.**:**  * :*:


seq1            IPTTHFHRGSASEGRAVLTSKTVKDFMLQKLNSLDIKGNA
seq2            IPTTHFHRGSASEGRAVLTSKTVKDFMLQKLNSLDIKGNA
seq3            IPTTHFHRGNASEGRAVLTSKTVKDFMLQKLNSLDIKGNA
seq4            SSATHFHRGSASEGRAVLTNKVVKDFMLQTLNDIDIRGSA
seq5            SSATHFHRGSASEGRAVLTNKVVKDFMLQTLNDIDIRGSA
                 .:******.********.*.*******.**.:**:*.*
```

Phylip format

```
 5 100
seq1         VANITLSTQH YRIHRSDVEP VKEKTTDKDV FAKSITAVRN SFISLSTSLS DRFSLHLQTD
seq2         VTNITLSTQH YRIHRSDVEP VKEKTTEKDI FAKSITAVRN SFISLSTSLS DRFSLHQQTD
seq3         VTNITLSTQH YRIHRSDVEP VKEKTTEKDI FAKSITAVRN SFISLSTSLS DRFSLHQQTD
seq4         VTKITLSPQN FRIQKQET-- LKEKSTEKNS LAKSILAVKN HFIELRSKLS ERFISHKNTE
seq5         VTKITLSPQN FRIQKQETTL LKEKSTEKNS LAKSILAVKN HFIELRSKLS ERFISHKNTE

             IPTTHFHRGS ASEGRAVLTS KTVKDFMLQK LNSLDIKGNA
             IPTTHFHRGS ASEGRAVLTS KTVKDFMLQK LNSLDIKGNA
             IPTTHFHRGN ASEGRAVLTS KTVKDFMLQK LNSLDIKGNA
             SSATHFHRGS ASEGRAVLTN KVVKDFMLQT LNDIDIRGSA
             SSATHFHRGS ASEGRAVLTN KVVKDFMLQT LNDIDIRGSA
```

Parse Multiple Sequence Alignment Results

1. Slice part of the alignment; 2. change format

script5.pl

```
#!/usr/local/bin/perl
use strict;
use warnings;

use Bio::AlignIO;

my $in = Bio::AlignIO->new(-file   => "myalignment.aln",
                           -format => "clustalw" );

my $out = Bio::AlignIO->new(-file   => ">out.phylip" ,
                            -format => 'phylip');

while ( my $aln = $in->next_aln() ) {
        my $new_aln = $aln->slice(5,100);
        $out->write_aln($new_aln);
}
```

Methods Implemented in Bio::SimpleAlign

- Modifier methods
  - add_seq
  - remove_seq
  - purge
  - sort_alphabetically
  - sort_by_list
  - set_new_reference
  - uniq_seq
- Sequence selection methods
  - each_seq
  - each_alphabetically
  - each_seq_with_id
  - get_seq_by_pos
  - get_seq_by_id
  - seq_with_features
- Create new alignments
  - select
  - select_noncont
  - slice
  - remove_columns
  - remove_gaps
- Change sequences within the MSA
  - splice_by_seq_pos
  - map_chars
  - uppercase
  - cigar_line
  - match_line
  - gap_line
  - all_gap_line
  - gap_col_matrix
  - match
  - unmatch
- MSA attributes
  - id
  - accession
  - description
  - missing_char
  - match_char
  - gap_char
  - symbol_chars
- Alignment descriptors
  - score
  - consensus_string
  - consensus_iupac
  - consensus_meta
  - is_flush
  - length
  - maxdisplayname_length
  - max_metaname_length
  - num_residues
  - num_sequences
  - average_percentage_identity
  - percentage_identity
  - overall_percentage_identity
- Alignment positions
  - column_from_residue_numbe
- Sequence names
  - displayname
  - set_displayname_count
  - set_displayname_flat
  - set_displayname_normal
  - source

Exercise 1. Retrieve an E. coli genome from NCBI (Genbank accession NC_000913). Make a fasta file with 500bp upstream regions of all transcripts.
Hint: You can do this by modifying script1.pl of this lecture.

Exercise 2. Modify script4.pl, so that this script can take in a third parameter maximum evalue, and only output HSP with evalue below the cutoff.