BioPerl II

CSHL Programming for Biology 2008

```
       opt      E()
< 20   323      0:==
  22     0      0:            one = represents 184 library sequences
  24     2      0:=
  26    12      2:*
  28    61     26:*
  30   211    157:*=
  32   664    607:===*
  34  1779   1645:========*=
  36  3558   3379:==================*=
  38  5908   5584:==============================*==
  40  8049   7790:=========================================*=
  42 10001   9522:==================================================*===
  44 10660  10503:=========================================================*
  46 10987  10698:=========================================================*=
  48 10332  10242:========================================================*=
  50  9053   9346:=================================================*
  52  7736   8217:========================================== *
  54  6828   7018:====================================*
  56  5448   5863:============================== *
  58  4484   4813:======================== *
  60  3818   3899:====================*
  62  2942   3126:================*
  64  2407   2486:=============*
  66  1866   1965:==========*
  68  1495   1545:========*
  70  1169   1211:======*
  72   886    946:=====*
  74   708    738:====*
  76   542    574:===*
  78   451    446:==*
  80   355    347:=*
  82   271    265:=*
  84   211    210:=*
  86   151    163:*
  88   104    126:*          inset = represents 3 library sequences
  90   101     97:*
  92    78     75:*         :========================*=
  94    56     58:*         :====================*
  96    38     45:*         :============= *
  98    26     35:*         :=========  *
 100    26     27:*         :========*
 102    20     21:*         :=======*
 104    13     16:*         :======*
 106    22     12:*         :===*====
 108    10     10:*         :===*
 110     5      7:*         :==*
 112     4      6:*         :=*
 114     4      4:*         :=*
 116     3      3:*         :*
 118     9      3:*         :*==
>120   110      2:*         :*===================================
```

Sequence Database Searching

A Detailed look at BLAST parsing

- 3 Components
  - Result: Bio::Search::Result::ResultI
  - Hit: Bio::Search::Hit::HitI
  - HSP: Bio::Search::HSP::HSPI

BLAST Report

```
BLASTP 2.0MP-WashU [04-May-2006] [macosx-10.4-i386-ILP32F64 2006-05-08T00:02:58]

Copyright (C) 1996-2006 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference:  Gish, W. (1996-2006) http://blast.wustl.edu

Query=  BOSS_DROME Protein bride of sevenless precursor.
        (896 letters)

Database:  wormpep190
           23,771 sequences; 10,449,259 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                      Smallest
                                                                      Sum
                                                               High   Probability
Sequences producing High-scoring Segment Pairs:                Score  P(N)       N

F35H10.10       CE24945 WBGene00018073  status:Confirmed        182   1.6e-09    2
Y43F8C.16       CE34118 WBGene00012837  status:Partially_confirm...     88   0.12       1
M02H5.2 CE25951 WBGene00019740  locus:srt-31    7TM chemorece...        86   0.16       1
[Some Seqs removed]

>F35H10.10      CE24945 WBGene00018073  status:Confirmed        UniProt:Q20073_CAEEL
           protein_id:AAA81683.2
         Length = 1404

 Score = 182 (69.1 bits), Expect = 1.6e-09, Sum P(2) = 1.6e-09
 Identities = 75/315 (23%), Positives = 149/315 (47%)

Query:   511 YPFLFDGESVMFWRIKMDTWVATGLTAAILGLIATLAILVFIVVRISLGDVFEGNPTTSI 570
             Y  +F+  +   WR     +V   L   ++  +  +A+LV ++V++ L  V +GN +  I
Sbjct:  1006 YQSVFEHITTGHWRDHPHNYVLLALITVLV--VVAIAVLVLVLVKLYLR-VVKGNQSLGI 1062

Query:   571 LLLLSLILVFCSFVPYSIEYVGEQRNSHVTFEDAQTLNTLCAVRVFIMTLVYCFVFSLLL 630
```

BLAST Parsing Script

```
#!/usr/bin/perl -w
use strict;
use Bio::SearchIO;
my $cutoff = '0.001';
my $file = 'BOSS_DROME.CE_WUBLASTP';
my $in = Bio::SearchIO->new(-format => 'blast',
                            -file    => $file);
while( my $r = $in->next_result ) {
  print "Query is: ", $r->query_name, " ",
  $r->query_description," ",$r->query_length," aa\n";
  print " Matrix was ", $r->get_parameter('matrix'), "\n";
  while( my $h = $r->next_hit ) {
  last if $h->significance > $cutoff;
  print "Hit is ", $h->name, "\n";
  while( my $hsp = $h->next_hsp ) {
   print " HSP Len is ", $hsp->length('total'), " ",
         " E-value is ", $hsp->evalue, " Bit score ",
         $hsp->score, " \n",
         " Query loc: ",$hsp->query->start, " ",
         $hsp->query->end," ",
         " Sbject loc: ",$hsp->hit->start, " ",
         $hsp->hit->end,"\n";
  }
}
```

parse_blast.pl

BLAST Script Results

parse_blast.pl

```
Query is: BOSS_DROME Protein bride of sevenless precursor. 896 aa
 Matrix was BLOSUM62
Hit is F35H10.10
 HSP Len is 315  E-value is 1.6e-09 Bit score 182
 Query loc: 511 813  Sbject loc: 1006 1298
 HSP Len is 28  E-value is 1.6e-09 Bit score 39
 Query loc: 508 535  Sbject loc: 427 454
```

FASTA Report

```
# fasta35 -E 1 -H BOSS_DROME.fa wormpep190
FASTA searches a protein or DNA sequence data bank
 version 35.04 Aug. 28, 2008
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query: BOSS_DROME.fa
  1>>>BOSS_DROME Protein bride of sevenless precursor. - 896 aa
Library: wormpep190 10449259 residues in 23771 sequences

10449259 residues in 23771 sequences
Statistics:  Expectation_n fit: rho(ln(x))= 5.6522+/-0.000443; mu= 12.7762+/- 0.026
 mean_var=109.3856+/-23.016, 0's: 1 Z-trim: 2  B-trim: 0 in 0/59
 Lambda= 0.122629
Algorithm: FASTA (3.5 Sept 2006) [optimized]
Parameters: BL50 matrix (15:-5) ktup: 2
 join: 38, opt: 26, open/ext: -10/-2, width:  16
 Scan time:  2.980

The best scores are:                                      opt bits E(23771)
F35H10.10 CE24945 WBGene00018073 status:Confirmed  (1404)  207 48.2 9.2e-05

>>F35H10.10 CE24945 WBGene00018073 status:Confirmed UniP  (1404 aa)
 initn:  94 init1:  94 opt: 207  Z-score: 196.5  bits: 48.2 E(): 9.2e-05
Smith-Waterman score: 275; 22.7% identity (52.3% similar) in 728 aa overlap (207-847:640-1330)

        180       190       200       210       220           230
BOSS_D RAISIDNASLAENLLIQEVQFLQQCTTYSMGIFVDWELYKQLESVIKD---LEYNIWPIP
                                    : : . .   .: :: ..   :. ..  .:
F35H10 NQAGRNITIVPKSVFGYASALHGDSQESLKGYFSSGDTDASLVSVDSEHSALQRSFTALP
     610       620       630       640       650       660

           240       250       260       270       280
BOSS_D GTRAHLFPKVAHLLHQMPWGEKIASV-EIATETLEMYNEFMEAARQEHMCLM--------
```

FASTA Parsing Script

```
#!/usr/bin/perl -w
use strict;
use Bio::SearchIO;
my $cutoff = '0.001';
my $file = 'BOSS_DROME.CE_FASTP';
my $in = Bio::SearchIO->new(-format => 'fasta',
                            -file    => $file);
while( my $r = $in->next_result ) {
  print "Query is: ", $r->query_name, " ",
  $r->query_description," ",$r->query_length," aa\n";
  print " Matrix was ", $r->get_parameter('matrix'), "\n";
  while( my $h = $r->next_hit ) {
    last if $h->significance > $cutoff;
    print "Hit is ", $h->name, "\n";
    while( my $hsp = $h->next_hsp ) {
     print " HSP Len is ", $hsp->length('total'), " ",
           " E-value is ", $hsp->evalue, " Bit score ",
           $hsp->score, " \n",
           " Query loc: ",$hsp->query->start, " ",
           $hsp->query->end," ",
           " Sbject loc: ",$hsp->hit->start, " ",
           $hsp->hit->end,"\n";
    }
  }
}
```

FASTA Script Results

```
Query is: BOSS_DROME Protein bride of sevenless precursor. 896 aa
 Matrix was BL50  Method was FASTA
Hit is F35H10.10
 HSP Len is 728  E-value is 9.2e-05 Bit score 196.5
 Query loc: 207 847  Sbject loc: 640 1330
```

Using the Search::Result object

```
use Bio::SearchIO;
use strict;
my $parser = new Bio::SearchIO(-format => 'blast',
                               -file => 'file.bls');
while( my $result = $parser->next_result ){
  print "query name=", $result->query_name, " desc=",
        $result->query_description, ", len=",$result->query_length,"\n";
  print "algorithm=", $result->algorithm, "\n";
  print "db name=", $result->database_name, " #lets=",
  $result->database_letters, " #seqs=",$result->database_entries, "\n";
  print "available params  ", join(',',
        $result->available_parameters),"\n";
  print "available stats ", join(',',
        $result->available_statistics), "\n";
  print "num of hits ", $result->num_hits, "\n";
}
```

Using the Search::Hit Object

```
use Bio::SearchIO;
use strict;
my $parser = new Bio::SearchIO(-format => 'blast',
                               -file => 'file.bls');
while( my $result = $parser->next_result ){
  while( my $hit = $result->next_hit ) {
   print "hit name=",$hit->name, " desc=", $hit->description,
         "\n len=", $hit->length, " acc=", $hit->accession, "\n";
   print "raw score ", $hit->raw_score, " bits ", $hit->bits,
         " significance/evalue=", $hit->evalue, "\n";
  }
}
```

Cool Hit Methods

- start(), end() - get overall alignment start and end for all HSPs
- strand() - get best overall alignment strand
- matches() - get total number of matches across entire set of HSPs (can specify only exact ‘id’ or conservative ‘cons’)

Using the Search::HSP Object

```
use Bio::SearchIO;
use strict;
my $parser = new Bio::SearchIO(-format => 'blast', -file => 'file.bls');
while( my $result = $parser->next_result ){
  while( my $hit = $result->next_hit ) {
     while( my $hsp = $hit->next_hsp ) {
       print "hsp evalue=", $hsp->evalue, " score=", $hsp->score, "\n";
       print "total length=", $hsp->hsp_length, " qlen=",
             $hsp->query->length, " hlen=",$hsp->hit->length, "\n";
       print "qstart=",$hsp->query->start, " qend=",$hsp->query->end,
             " qstrand=", $hsp->query->strand, "\n";
       print "hstart=",$hsp->hit->start, " hend=",$hsp->hit->end,
             " hstrand=", $hsp->hit->strand, "\n";
       print "percent identical ", $hsp->percent_identity,
             " frac conserved ", $hsp->frac_conserved(), "\n";
       print "num query gaps ", $hsp->gaps('query'), "\n";
       print "hit str =", $hsp->hit_string, "\n";
       print "query str =", $hsp->query_string, "\n";
       print "homolog str=", $hsp->homology_string, "\n";
  }
}
```

Cool HSP methods

- rank() - order in the alignment (which you could have requested, by score, size)
- matches - overall number of matches
- seq_inds - get a list of numbers representing residue positions which are
  - conserved, identical, mismatches, gaps

SearchIO system

- BLAST (WU-BLAST, NCBI, XML, PSIBLAST, BL2SEQ, MEGABLAST, TABULAR (-m8/m9))
- FASTA (m9 and m0)
- HMMER (hmmpfam, hmmsearch)
- UCSC formats (WABA, AXT, PSL)
- Gene based alignments
  - Exonerate, SIM4, {Gene,Genome}wise

SearchIO reformatting

- Supports output of Search reports as well
- Bio::SearchIO::Writer
  - "BLAST flavor" HTML, Text
  - Tabular Report Format

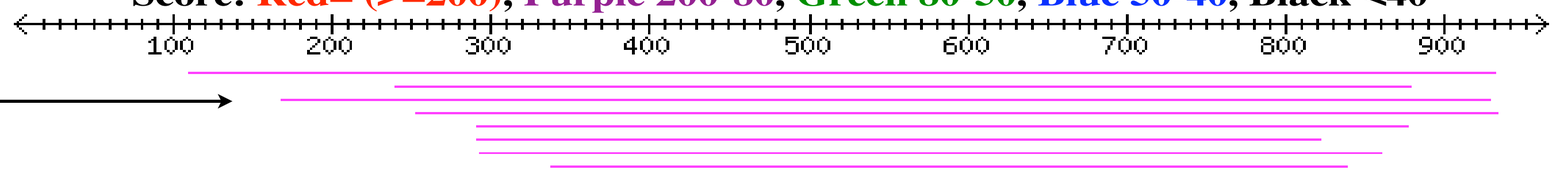
Bioperl Reformatted HTML of BLASTP Search Report for gi|6319512|ref|NP_009594.1|

BLASTP 2.0MP-WashU [04-Feb-2003] [linux24-i686-ILP32F64 2003-02-04T19:05:09]

Reference: Gish, W. (1996-2000) http://blast.wustl.edu

Hit Overview

Score: Red= (>=200), Purple 200-80, Green 80-50, Blue 50-40, Black <40



Build custom overview (Bio::Graphics)

Query= gi|6319512|ref|NP_009594.1| chitin synthase 2; Chs2p [Saccharomyces cerevisiae]
(963 letters)

Database: cneoA_WI.aa
9,645 sequences; 2,832,832 total letters

Hyperlink to external resources

Sequences producing significant alignments:	Score (bits)	E value
cneo_WIH99_157.Gene2 Start=295 End=4301 Strand=1 Length=912 ExonCt=24	1650	1.6e-173
cneo_WIH99_63.Gene181 Start=154896 End=151527 Strand=-1 Length=876 ExonCt=13	1441	3.9e-149
cneo_WIH99_133.Gene1 Start=15489 End=19943 Strand=1 Length=1017 ExonCt=23	1357	3e-142
cneo_WIH99_45.Gene2 Start=84 End=3840 Strand=1 Length=839 ExonCt=25	1311	1.5e-138
cneo_WIH99_112.Gene165 Start=122440 End=118921 Strand=-1 Length=1036 ExonCt=9	198	1.2e-15
cneo_WIH99_11.Gene7 Start=39355 End=42071 Strand=1 Length=761 ExonCt=9	172	6.4e-13
cneo_WIH99_60.Gene9 Start=36153 End=32819 Strand=-1 Length=1020 ExonCt=5	166	1.2e-12
cneo_WIH99_106.Gene88 Start=242538 End=238790 Strand=-1 Length=1224 ExonCt=3	157	6.3e-09

Hyperlink to alignment part of report

Turning BLAST into HTML

```
use Bio::SearchIO;
use Bio::SearchIO::Writer::HTMLResultWriter;

my $in = new Bio::SearchIO(-format => 'blast'',
           -file   => shift @ARGV);

  my $writer = new Bio::SearchIO::Writer::HTMLResultWriter();
  my $out = new Bio::SearchIO(-writer => $writer
                              -file    => ">file.html");
  $out->write_result($in->next_result);
```

Turning BLAST into HTML

```
# to filter your output
  my $MinLength = 100; # need a variable with scope outside the method
  sub hsp_filter {
      my $hsp = shift;
      return 1 if $hsp->length('total') > $MinLength;
  }
  sub result_filter {
      my $result = shift;
      return $hsp->num_hits > 0;
  }

  my $writer = new Bio::SearchIO::Writer::HTMLResultWriter
                       (-filters => { 'HSP' => \&hsp_filter} );
  my $out = new Bio::SearchIO(-writer => $writer);
  $out->write_result($in->next_result);

  # can also set the filter via the writer object
  $writer->filter('RESULT', \&result_filter);
```

Multiple Alignments

- Bio::AlignIO for parsing and writing Multiple Alignments file formats including
  - phylip, nexus,clustalw, msf, mega, meme,pfam,psi,selex,stockholm
- Parser produces Bio::SimpleAlign objects
  - To extract or remove columns
  - Calculate consensus string and percent identity

Alignments format conversion

```
#!/usr/bin/perl -w
use strict;
use Bio::AlignIO;
my $in = Bio::AlignIO->new(-format => 'fasta',
                           -file   => 'actin_hits.fasaln');
my $out = Bio::AlignIO->new(-format => 'nexus',
                            -file   => '>actin_hits.nex');
while( my $aln = $in->next_aln ){
 $out->write_aln($aln);
}
```

View a piece of alignment

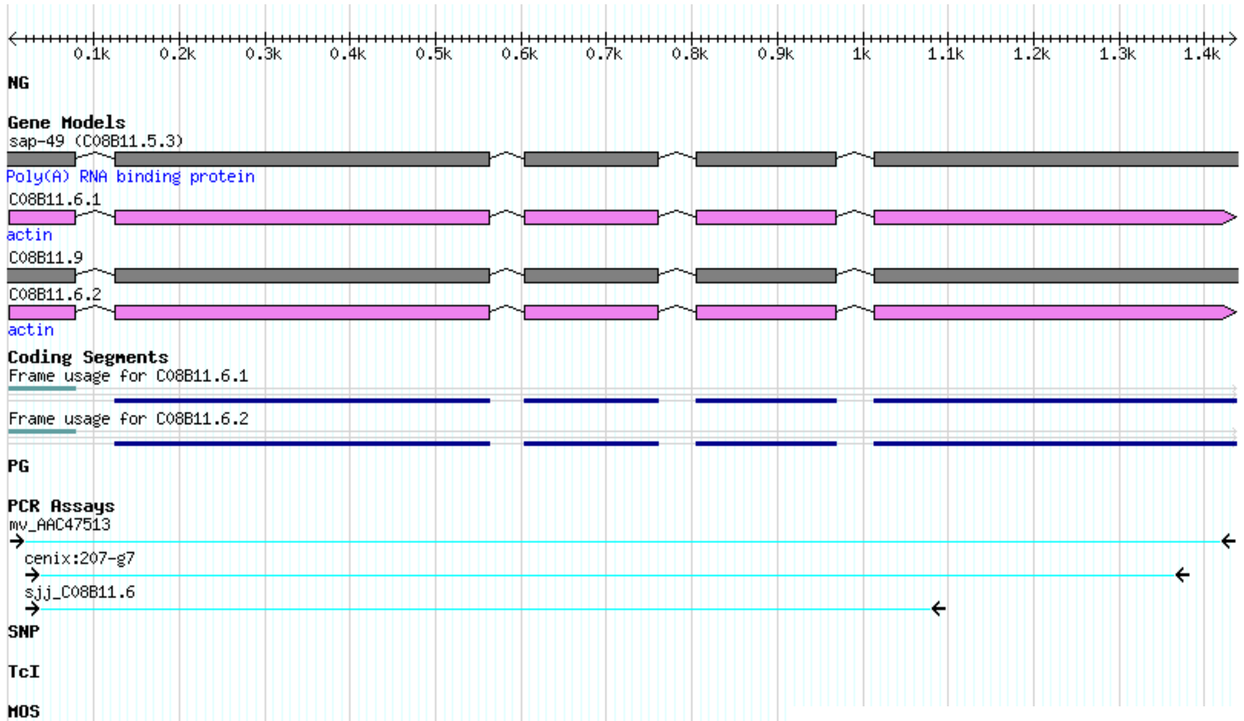
```
#NEXUS
[TITLE: NoName]

begin data;
dimensions ntax=50 nchar=1706;
format interleave datatype=protein   gap=- symbols="SFTNKEYVQMCLAWPHDIRG";

matrix
umay_BRD_UM03743_1                    ---------- ---------- ---------- ---------- ----------
rory_SNAP_rory_1_3_snap_337           ---------- ---------- ---------- ---------- ----------
umay_BRD_UM05510_1                    MIFVVVHTNR RLFTDPLPAA RPARLKLSKH KTQTRTSRSS AIMQTATTSN
cneo_WM276_GLEAN_GLEAN_03860          ---------- ---------- ---------- ---------- ----------
cneo_JEC21_TIGR_CNK01830              ---------- ---------- ---------- ---------- ----------
pchr_GLEAN_GLEAN_gz_02127             ---------- ---------- ---------- ---------- ----------
ccin_GLEAN_GLEAN_gz2_09189            ---------- ---------- ---------- ---------- ----------
rory_SNAP_rory_1_65_snap_42           ---------- ---------- ---------- ---------- ----------
rory_SNAP_rory_1_69_snap_40           ---------- ---------- ---------- ---------- ----------
rory_SNAP_rory_1_25_snap_254          ---------- ---------- ---------- ---------- ----------
cneo_JEC21_TIGR_CND05140              ---------- ---------- ---------- ---------- ----------
cneo_H99_GLEAN_GLEAN_02086            ---------- ---------- ---------- ---------- ----------
cneo_WM276_GLEAN_GLEAN_04292          ---------- ---------- ---------- ---------- ----------
umay_BRD_UM01034_1                    ---------- ---------- ---------- ---------- ----------
ccin_GLEAN_GLEAN_gz2_01080            ---------- ---------- ---------- ---------- ----------
pchr_GLEAN_GLEAN_gz_10587             ---------- ---------- ---------- ---------- ----------
rory_SNAP_rory_1_2_snap_334           ---------- ---------- ---------- ---------- ----------
rory_SNAP_rory_1_55_snap_33           ---------- ---------- ---------- ---------- ----------
```

$ more actin_hits.msf

Alignment manipulation

- Remove some sequences and rewrite the result
- Extract the columns that correspond to a motif found in one particular sequence

Worm gene's structure



Predicted Exon Structure

NOTE: Transcript is on (-) strand.

	Relative to Itself		Relative to C08B11	
Exon #	Start	End	Start	End
1	1	77	21223	21147
2	125	562	21099	20662
3	605	760	20619	20464
4	806	967	20418	20257
5	1013	1436	20211	19788

Protein Report for: WP:CE30856

Close Window

Motif Details		
Feature	Start	End
interpro IPR004000	3	389

http://wormbase.org/db/seq/sequence?name=C08B11.6.2;class=Transcript

Alignment manipulation

```
#!/usr/bin/perl -w
use strict;
use Bio::AlignIO;
my %seqs_to_remove =(umay_BRD_UM05510_1 => 1, umay_BRD_UM01600_1 =>1);

my $in = Bio::AlignIO->new(-format => 'fasta',
                           -file   => 'actin_hits.fasaln');
my $out = Bio::AlignIO->new(-format => 'nexus',
                            -file   => '>actin_hits_trim.nex');
if( my $aln = $in->next_aln ){
 my @seqs = $aln->each_seq;
 for my $s ( @seqs ) {
   if( exists $seqs_to_remove{$s->id} ) {
      $aln->remove_seq($s);
   }
 }
 my $updated = $aln->remove_gaps('-',0); # only remove the 'all gap' cols
 $out->write_aln($updated);
}
```

$ perl alignment_manip.pl

```
#NEXUS
[TITLE: NoName]

begin data;
dimensions ntax=48 nchar=1142;
format interleave datatype=protein   gap=- symbols="SFTNKEYVQMCLAWPHDIRG";

matrix
umay_BRD_UM03743_1                      ---------- ---------M AAASFSQSAE AIDPLAGVDY IIVIDNGAHN
rory_SNAP_rory_1_3_snap_337             MVEASI--P- -RYYSLEEKE YSAPYYNIRN DYKTFHSLKT PIVIDNGSKQ
cneo_WM276_GLEAN_GLEAN_03860            MSGNLIDIPE LRLN-----E EPQPV----F DYHSLDGQSS AICIDNGAYS
cneo_JEC21_TIGR_CNK01830                MPDNLIDIPE IRFN-----G EPQPV----F DYHSLDGQSP AICIDNGAHS
pchr_GLEAN_GLEAN_gz_02127               -MAETPELI- -RIP-----N PPLPTVRQPA SYDEFRGTGT PLIIDNGSTN
ccin_GLEAN_GLEAN_gz2_09189              -MQNTFHLP- -IYT-----P PSLPI--QAE SYDVHRENGT PLIIDNGATT
rory_SNAP_rory_1_65_snap_42             ---------- ---------- ---------- MVTYGGDEVN AIVMDMGSTS
rory_SNAP_rory_1_69_snap_40             ---------- ---------- ---------- ---------- -----MGTCS
rory_SNAP_rory_1_25_snap_254            ---------- ---------- ---------- -----MTKLP VVVMDNGTGY
cneo_JEC21_TIGR_CND05140                ---------- ---------- ---------- -----MSRQP PLVIDNGTGY
cneo_H99_GLEAN_GLEAN_02086              ---------- ---------- ---------- -----MSRQP PLVIDNGTGY
cneo_WM276_GLEAN_GLEAN_04292            ---------- ---------- ---------- -----MSRQP PLVIDNGTGY
umay_BRD_UM01034_1                      ---------- ---------- ---------- -----MSRSN VIVLDNGTGY
ccin_GLEAN_GLEAN_gz2_01080              ---------- ---------- ---------- -----MAYLA PIISDNGTGY
pchr_GLEAN_GLEAN_gz_10587               ---------- ---------- ---------- -----MSLLA PIICDNGTGF
rory_SNAP_rory_1_2_snap_334             ---------- ---------- ---------M YNPSVCLKST YIIVDNGSKF
rory_SNAP_rory_1_55_snap_33             ---------- ---------- ---------- -----MNASK TLVVDNGTGF
rory_SNAP_rory_1_5_snap_483             ---------- ---------- ---------- -----MTTSK TLVVDNGTGF
umay_BRD_UM05405_1                      ---------- ---------- ---------- -----MADQR PVVVDNGTGF
```

Mapping between coordinates

```
#!/usr/bin/perl -w
use strict;
use Bio::AlignIO;
my $ref_seq = 'C08B11_6';
my $pos_start = 3; #INTERPRO ACTIN DOMAIN START
my $pos_end   = 389; #INTERPRO ACTIN DOMAIN START
my $in = Bio::AlignIO->new(-format => 'fasta',
                           -file   => 'actin_hits.fasaln');
my $out = Bio::AlignIO->new(-format => 'nexus',
                            -file   => '>actin_hits_domain.nex');
if( my $aln = $in->next_aln ){
 for my $s ($aln->each_seq ) {
   if(  $s->id eq $ref_seq ) {
     my $col_start = $aln->column_from_residue_number($s->id, $pos_start);
     my $col_end = $aln->column_from_residue_number($s->id, $pos_end);
     print "grabbing columns $col_start .. $col_end\n";
     my $piece = $aln->slice($col_start, $col_end);
     $out->write_aln($piece);
     last; # all done
   }
 }
}
```

Databases of Annotations & Features

- Bio::DB::GFF
- Bio::DB::SeqFeature
- BioSQL
- Other APIs and Relational Databases for Sequence & features
  - MODWare + Chado

Feature overlap test

How Many (or which) SNPs are in Genes

```
#!/usr/bin/perl -w
use strict;
use Bio::DB::GFF;
my $dbh = Bio::DB::GFF->new(-adaptor => 'dbi::mysqlopt',
                            -user => 'USER', -pass=>'PASSWORD'
                            -dsn => 'dbi:mysql:database=DBNAME');
my $iterator = $dbh->get_seq_stream(-type     => 'mRNA');
my $total_genic_SNPs;
while (my $s = $iterator->next_seq) {
 # don't care about strand since SNPs are single bp
 my $segment = $dbh->segment($s->seq_id,$s->start => $s->end);
 my @snps = $segment->features('SNP');
 $total_genic_SNPs += scalar @snps;
}
```

Other things

- Molecular evolution tools - PAML parsing & running, simple molecular distances
- PopGen - summary pop stats (theta, pi, Tajima's D), simple LD
- Links to slides from last year on course site

BioPerl Pipeline building

- Whole set of wrappers for running Bioinformatics tools in bioperl-run
- Run BLAST locally or sub remote jobs (through NCBI)
- Run PAML - handles setup and take down of temporary files and directories
- Run alignment progs through similar interfaces - T-Coffee, MUSCLE, Clustalw

Thanks

- It is a collaborative project among many different individuals. Credits are given on the website.
  - IRC and Twitter and Blogs can be found linked through website too.
- You too can contribute! Join mailing list or read it on the web.